TOPICS IN SEQUENCE ANALYSIS

A Thesis
Presented to
The Academic Faculty

by

Jinyong Ma

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Mathematics

Georgia Institute of Technology
December 2012

TOPICS IN SEQUENCE ANALYSIS

Approved by:

Professor Christian Houdré, Advisor
School of Mathematics
Georgia Institute of Technology

Professor Yuri Bakhtin
School of Mathematics
Georgia Institute of Technology

Professor Robert D. Foley
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Vladimir Koltchinskii
School of Mathematics
Georgia Institute of Technology

Professor Heinrich Matzinger
School of Mathematics
Georgia Institute of Technology

Professor Ionel Popescu
School of Mathematics
Georgia Institute of Technology

Date Approved: 5 November 2012

To my family.

ACKNOWLEDGEMENTS

I first would like to thank my advisor, Dr. Christian Houdré. Not only did he teach me Mathematics and inspire my research, he also helped me navigate issues of both life and study from the very first day I arrived at Georgia Tech. I would be always grateful for his patient and valuable guidance on every aspect of my work.

I also would like to thank Dr. Ionel Popescu, Dr. Vladimir Koltchinskii, Dr. Yuri Bakhtin, Dr. Xingxing Yu, Dr. Michael Loss, among others, for teaching me mathematics and encouraging me with my research. In addition, I am thankful for Dr. Ionel Popescu, Dr. Vladimir Koltchinskii, Dr. Yuri Bakhtin, Dr. Robert D. Foley and Dr. Heinrich Matzinger for serving on my dissertation committee and providing helpful comments and feedback on my dissertation. I would like to thank Dr. Ronghua Pan, who introduced me to come to Georgia Tech and helped me a lot in my study and life.

I would also like to extend my sincere gratitude to Dr. Luca Dieci, who extended every effort to help me get used to the new Georgia Tech environment and make everything smooth for my study, and to Dr. John Etnyre, who kindly continued to help me to get the work done. I particularly want to thank Ms. Cathy Jacobson, who not only gave the best English training program to make my teaching easy from the beginning, but also enhanced my life here in every possible way. My special thanks goes to Ms. Klara Grodzinsky, for her gracious support of my teaching, and to Ms. Sharon McDowell, Ms. Karen Hinds, Ms. Genola Turner and the IT group for their everyday support.

There are many of my fellow graduate students whose friendships have meant so much over the years. I especially want to thank to Hua Xu, Jing Xu and Luan

Lin, who helped me settle down when I first came to Georgia Tech. I have enjoyed tremendous happiness and shared learning experiences with all my friends, notably Hao Deng, Ruoting Gong, Allen Hoffmeyer, Liangda Huang, Xun Huang, Huy Huynh, Xiaxi Li, Yao Li, Yongfeng Li, Nan Lu, Kai Ni, Bohuan Wei, Yuejian Xie, Tianjun Ye, Ke Yin, Jia Zeng, Weizhe Zhang.

I owe so much to my parents who have been educating and supporting me in all of my lifes choices. My special thanks goes to my two sisters and my two lovely nephews, for their support and love. Last but not least, my heartfelt gratitude goes to my girlfriend, Jing, for her continuous backing and faithful love.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

This thesis studies two topics in sequence analysis. In the first part, we investigate the large deviations of the shape of the random RSK Young diagrams, associated with a random word of size n whose letters are independently drawn from an alphabet of size $m = m(n)$. When the letters are drawn uniformly and when both n and m converge together to infinity, m not growing too fast with respect to n, the large deviations of the shape of the Young diagrams are shown to be the same as that of the spectrum of the traceless GUE. Since the length of the top row of the Young diagrams is the length of the longest (weakly) increasing subsequence of the random word, the corresponding large deviations follow. When the letters are drawn with non-uniform probability, a control of both highest probabilities will ensure that the length of the top row of the diagrams satisfies a large deviation principle. In either case, both speeds and rate functions are identified. To complete our study, non-asymptotic concentration bounds for the length of the top row of the diagrams, are obtained for both models.

In the second part, we investigate the order of the r-th, $1 \leq r < +\infty$, central moment of the length of the longest common subsequence of two independent random words of size n whose letters are identically distributed and independently drawn from a finite alphabet. When all but one of the letters are drawn with small probabilities, which depend on the size of the alphabet, the r-th central moment is shown to be of order $n^{r/2}$. In particular, when $r = 2$, we get the order of the variance of the longest common subsequence.

CHAPTER I

SIMULTANEOUS LARGE DEVIATIONS FOR THE SHAPE OF YOUNG DIAGRAMS ASSOCIATED WITH RANDOM WORDS

1.1 Introduction and results

Let $\mathcal{A}_m = \{\alpha_1 < \alpha_2 < \cdots < \alpha_m\}$ be an ordered alphabet of size m, and let a word be made of the random letters $X_1^m, X_2^m, \cdots, X_n^m$, independently drawn from $\mathcal{A}_m$. The Robinson-Schensted-Knuth (RSK) correspondence associates to this random word a pair of Young diagrams, of the same shape, having at most m rows. Now for $i = 1, 2, \cdots, m$, let $R_i(n, m)$ denote the length of the ith row of the Young diagrams, and recall that $R_1(n, m)$, the length of the top row, coincides with the length of the longest increasing subsequence of the random word $X_1^m X_2^m \cdots X_n^m$. Let us take the sequence (4,1,3,4,2) as example, the RSK correspondence is shown in Figure 1, with the shape of the Young diagrams is $(R_1, R_2, R_3) = (3, 1, 1)$.

Appropriately renormalized and for uniform draws, the shape $(R_i(n, m))_{i=1}^m$ of the Young diagrams converges, in law, to the spectrum of an $m \times m$ element of the traceless GUE ([25], [39]). In turn, any fixed size subset of this spectrum, also

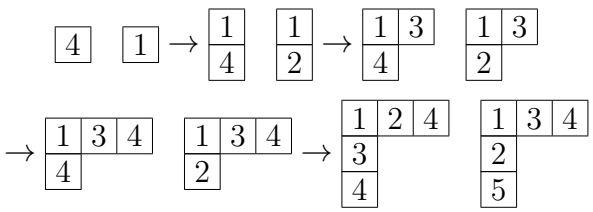

Figure 1: RSK correspondence for the sequence (4,1,3,4,2).

converges with m, and after proper renormalization, to a multidimensional Tracy-Widom distribution ([38], [40]). These asymptotics have further led (see [11]) to the study of the limiting shape of these Young diagrams when the word length and alphabet size simultaneously grow to infinity. This is briefly recalled next.

Let the random matrix $\mathbf{X} = (\mathbf{X}_{ij})_{1\le i,j\le m}$ be an element of the $m \times m$ GUE with rescaling such that $Re(\mathbf{X}_{ij}) \sim N(0, 1/2)$ and $Im(\mathbf{X}_{ij}) \sim N(0, 1/2)$, for $i \neq j$; and $\mathbf{X}_{ii} \sim N(0, 1)$ (see [5] and [32] for background on random matrices). Let $(\lambda_1^m, \lambda_2^m, \cdots, \lambda_m^m)$ be the nonincreasing ordered spectrum of $\mathbf{X}$, and let $(\lambda_1^{m,0}, \lambda_2^{m,0}, \cdots, \lambda_m^{m,0})$ be the corresponding ordered spectrum of an element of the traceless GUE (that is of $\mathbf{X} - tr(\mathbf{X})/m$). An important fact (*e.g.* [8], [18], [20]) asserts that

$$\begin{aligned}&\left(\lambda_1^{m,0}, \lambda_2^{m,0}, \cdots, \lambda_m^{m,0}\right)\\ &\stackrel{\mathcal{L}}{=} \frac{\sqrt{m-1}}{\sqrt{m}}\mathbf{\Theta}_m^{-1}\left(\left(\max_{\mathbf{t}\in I_{k,m}}\sum_{j=1}^{k}\sum_{l=j}^{m-k+j}\left(\tilde{B}^l_{t_{j,l}} - \tilde{B}^l_{t_{j,l-1}}\right)\right)_{1\le k\le m}\right),\end{aligned} \tag{1.1.1}$$

where $\mathbf{\Theta}_m : \mathbb{R}^m \to \mathbb{R}^m$ is defined via $(\mathbf{\Theta}_m(\mathbf{x}))_j = \sum_{i=1}^j x_i, 1 \le j \le k$, and where $(\tilde{B}^j)_{1\le j\le m}$ is a driftless m-dimensional Brownian Motion with covariance matrix

$$t\begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}, \tag{1.1.2}$$

with $\rho = -1/(m-1)$, and where for $1 \le k \le m$,

$$\begin{aligned}I_{k,m} = \{&\mathbf{t} = (t_{j,l} : 1 \le j \le k, 0 \le l \le m) : t_{j,j-1} = 0, t_{j,m-k+j} = 1, 1 \le j \le k,\\ &t_{j,l-1} \le t_{j,l}, 1 \le j \le k, 1 \le l \le m-1; t_{j,l} \le t_{j-1,l-1}, 2 \le j \le k, 2 \le l \le m\}.\end{aligned}$$

By comparing the Brownian functionals in (1.1.1) with discrete functionals representing the shape of the Young diagrams, and via a KMT approximation, the simultaneous asymptotic convergence of the shape of the random RSK Young diagrams is obtained in [11].

A related strategy is pursued here in order to investigate the large deviations of the shape of the RSK Young diagrams. More precisely, we obtain a large deviation principle for the length of the first r rows of the Young diagrams, when n and m simultaneously converge to infinity and when the size m of the alphabet does not grow too fast. To achieve our goals, we also rely on the techniques and results developed in [9] (see also [4]), where large deviations are obtained for the largest (or the rth largest) eigenvalue of the GOE. These methodologies further give the multidimensional large deviations for the first r eigenvalues of the ordered spectrum of the traceless GUE. In turn, combined with a KMT approximation, these lead to large deviations for the shape of the diagrams.

Let us put our work into context. For random permutations, the large deviations of the length of the longest increasing subsequence are described in [17] and [36], while, moderate deviations are given in [30] and [31]. Closer to our framework, in [23], following the comparison method of [7] and [12], large deviations for the last-passage directed percolation model close to the x-axis are established for *iid* Gaussian or bounded weights. The length of the top row of the diagrams also corresponds to a last-passage percolation, but with *dependent* (exchangeable in the uniform case) Bernoulli weights (see (1.2.3)). In our framework, we also take care of the other rows of the diagrams.

Now recall that for a Polish space $\mathcal{X}$, the function $I : \mathcal{X} \to \mathbb{R}^+ \cup \{+\infty\}$ is called a rate function, if it is lower semi-continuous, i.e. if its level sets $\{x \in \mathcal{X} : I(x) \leq M\}$ are closed for any $M \geq 0$, and is called a good rate function if its level sets $\{x \in \mathcal{X} : I(x) \leq M\}$ are compact for any $M \geq 0$. Recall also that a sequence $(\mu_n)_{n\in\mathbb{N}}$ of probability measures on $\mathcal{X}$ satisfies a large deviation principle with speed (or in the scale) a_n (going to infinity with n) and rate function I if and only if for any closed subset F of $\mathcal{X}$,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(F) \leq -\inf_F I, \tag{1.1.3}$$

and for any open subset O of $\mathcal{X}$,

$$\liminf_{n\to\infty} \frac{1}{a_n} \log \mu_n(O) \geq -\inf_O I. \tag{1.1.4}$$

Here is one of the main results of our work,

Theorem 1.1.1 *In the uniform case, let m and n simultaneously converge to infinity in such a way that $m(n) = o(n^{1/4})$. Then, for any $r \geq 1$,*

$$\left(\frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}}, \cdots, \frac{R_r(n, m(n)) - n/m(n)}{\sqrt{n}} \right)$$

satisfies a large deviation principle with speed $m(n)$ and good rate function I_r on the space $\mathcal{L}^r := \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \geq x_2 \geq \cdots \geq x_r\}$, where

$$I_r(x_1, x_2, \cdots, x_r) = \begin{cases} 2\sum_{i=1}^r \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz, & \text{if } x_1 \geq x_2 \cdots \geq x_r \geq 2, \\ +\infty, & \text{otherwise.} \end{cases} \tag{1.1.5}$$

In other words, for all $x_1 \geq x_2 \cdots \geq x_r \geq 2$,

$$\lim_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\Bigg(\frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}} \geq x_1, \cdots, \frac{R_r(n, m(n)) - n/m(n)}{\sqrt{n}} \geq x_r \Bigg) = -2 \sum_{i=1}^r \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz, \tag{1.1.6}$$

while for any $x < 2$ and $1 \leq i \leq r$,

$$\lim_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{R_i(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x \right) = -\infty. \tag{1.1.7}$$

Remark 1.1.1 *The rate function I_r in* (1.1.5) *is a good rate function. Moreover it is continuous and increasing with respect to each individual variable on its effective domain $\mathcal{D}_{I_r} = \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \geq x_2 \geq \cdots \geq x_r \geq 2\}$, given that the other variables are fixed. Thus, when proving the large deviation principle (LDP) as in Theorem 1.1.1, instead of proving both the usual upper and lower bounds, i.e., that for any closed set F in $\mathcal{L}^r = \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \geq x_2 \geq \cdots \geq x_r\}$,*

$$\limsup_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(X_r^n \in F\right) \leq -\inf_F I_r, \tag{1.1.8}$$

and that for any open set O in $\mathcal{L}^r$,

$$\liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(X_r^n \in O\right) \geq -\inf_O I_r, \tag{1.1.9}$$

where

$$X_r^n = \left(\frac{R_i(n, m(n)) - n/m(n)}{\sqrt{n}} \right)_{1 \leq i \leq r},$$

it is enough to prove a limiting equality on rectangular subsets as in (1.1.6) *and* (1.1.7)*. The proof of this argument can be stated in this way:*

Recall that a cube in $\mathbb{R}^r$ is a Cartesian product of r closed intervals whose side lengths are all equal. For $k \in \mathbb{Z}^+$, let $\mathcal{Q}_k$ be the collection of cubes whose side length is 2^{-k} and whose vertices are in the lattice $(2^{-k}\mathbb{Z})^r$. (That is, $\prod_1^r [a_j, b_j] \in \mathcal{Q}_k$ iff $2^k a_j$ and $2^k b_j$ are integers and $b_j - a_j = 2^{-k}$ for all j.) Note that any two cubes in $\mathcal{Q}_k$ have disjoint interiors, and that the cubs in $\mathcal{Q}_{k+1}$ are obtained from the cubes in $\mathcal{Q}_k$ by bisecting the sides. For any Borel set $E \subset \mathbb{R}^r$, we define the inner and outer approximations to E by the grid of cubes $\mathcal{Q}_k$ to be

$$\underline{A}(E,k) = \bigcup \{Q \in \mathcal{Q}_k : Q \subset E\}, \quad \overline{A}(E,k) = \bigcup \{Q \in \mathcal{Q}_k : Q \cap E \neq \phi\}.$$

Notice that for any cube $Q \in \mathcal{Q}_k$ and $Q \subset \mathcal{D}_{I_r}$, let $(v_1, \cdots, v_r)$ be the vertex of Q that is nearest to the origin, then

$$\lim_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(X_r^n \in Q\right) = -I_r(v_1, \cdots, v_r). \tag{1.1.10}$$

Fixing any open subset O of $\mathcal{L}^r$, for any $\epsilon > 0$, we may find one point $v \in O$ such that $\inf_O I_r \leq I_r(v) \leq \inf_O I_r + \epsilon$. Since the rate function is infinity outside the rectangular effective domain $\mathcal{D}_{I_r}$, so without loss of generality, we may assume that this point v belongs to the interior of $\mathcal{D}_{I_r}$. Given this, we can find a large enough natural number k and some cube $Q \in \mathcal{Q}_k$ such that $v \in Q \subset \mathcal{D}_{I_r}$. So from (1.1.10)*,*

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(X_r^n \in O\right) &\geq \liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(X_r^n \in Q\right) \\ &= -I_r(v_1, \cdots, v_r) \geq -I_r(v) \geq -\inf_O I_r - \epsilon.\end{aligned}$$

Letting ϵ go to zero, we get (1.1.9).

On the other hand, we first take F as any compact subset of $\mathcal{L}^r$. From (1.1.7),

$$\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in F\right)=\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in F\cap\mathcal{D}_{I_r}\right),$$

thus we may just assume that $F\subset\mathcal{D}_{I_r}$. For any positive integer k, we have

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in F\right)&\le\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in\overline{A}(F,k)\right)\\&=\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in\overline{A}(F,k)\cap\mathcal{D}_{I_r}\right).\end{aligned}$$

Since for fixed k the number of the cubes (or the intersection parts of the cubes with $\mathcal{D}_{I_r}$) is finite, from (1.1.10) *we can get that*

$$\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in\overline{A}(F,k)\cap\mathcal{D}_{I_r}\right)=-I_r(v_{F,k})$$

where $v_{F,k}$ is the point in $\overline{A}(F,k)\cap\mathcal{D}_{I_r}$ where the rate function achieves its infimum. However, from the definition of $\overline{A}(F,k)$, in F we can find some point $w_{F,k}$ which lies in the same cube with $v_{F,k}$, in other words, the distance between $w_{F,k}$ and $v_{F,k}$ is at most $2^{-k}\sqrt{r}$. As a consequence, since F is compact, so I_r is uniformly continuous on $\overline{A}(F,k)\cap\mathcal{D}_{I_r}$, so for any $\epsilon>0$, as long as k is big enough:

$$\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(X_r^n\in F\right)\le-I_r(v_{F,k})\le-I_r(w_{F,k})+\epsilon\le-\inf_F I_r+\epsilon.$$

Again, by letting ϵ go to 0, we can prove (1.1.8) *for compact set F. For the general closed set F, we just choose some large cube, say $Q_0=\{x=(x_1,\cdots,x_r):\max|x_j|\le 2^M,1\le i\le r\}$, then consider $F\cap Q_0$ and $F\backslash Q_0$ separately, we may get* (1.1.8) *holds for any closed set F.*

In Theorem 1.1.1, if at least one of the renormalized variables is on the left of its simultaneous asymptotic mean, by changing the convergence speed from m to m^2, a more accurate form of (1.1.7) is valid. The closed form expression obtained for K below was found after Satya Majumdar kindly suggested that the methodology developed in [33] would apply to our traceless GUE framework.

Theorem 1.1.2 *In the uniform case, let m and n simultaneously converge to infinity in such a way that $m(n) = o(n^{1/6})$. Then, for any $r \geq 1$,*

$$\left(\frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}}, \cdots, \frac{R_r(n, m(n)) - n/m(n)}{\sqrt{n}} \right)$$

satisfies a large deviation principle with speed $(m(n))^2$ and good rate function $K(x_r)$ on the space $\mathcal{L}^r := \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \geq x_2 \geq \cdots \geq x_r\}$, where K is the rate function of the largest eigenvalue of the $m \times m$ traceless GUE, when on the left of its asymptotic mean. It is given by

$$K(x) := \inf_{\mu \in \mathcal{M}_0((-\infty, x])} I(\mu), \tag{1.1.11}$$

where I (see (1.5.5)) is the rate function for the LDP of the spectral measure of the GUE, and $\mathcal{M}_0((-\infty, x])$ is the set of zero mean probability measures supported on $(-\infty, x]$. For $x \leq 0$, $K(x) = +\infty$, for $x \geq 2$, $K(x) = 0$, and for $0 < x < 2$,

$$\begin{aligned} K(x) = \frac{1}{48} \Bigg(& 3 \left(9 \sqrt[3]{2} 23^{2/3} \left(\sqrt{81x^2 + 12} - 9x \right)^{2/3} - 8 \right) x^2 + \\ & 9 \sqrt[3]{2} \sqrt[6]{3} \left(\sqrt{81x^2 + 12} - 9x \right)^{1/3} \left(\sqrt{27x^2 + 4} \left(\sqrt{81x^2 + 12} - 9x \right)^{1/3} - 5 \sqrt[3]{2} \sqrt[6]{3} \right) x - \\ & 6 \sqrt[3]{2} 23^{2/3} \left(\sqrt{81x^2 + 12} - 9x \right)^{2/3} - 3 \, 2^{2/3} 3^{5/6} \sqrt{27x^2 + 4} \left(\sqrt{81x^2 + 12} - 9x \right)^{1/3} + \\ & 16 \log \left(\sqrt{81x^2 + 12} - 9x \right) - 48 \log \left(2 \sqrt[3]{3} - \sqrt[3]{2} \left(\sqrt{81x^2 + 12} - 9x \right)^{2/3} \right) + \\ & 60 + 32 \log 6 \Bigg). \end{aligned} \tag{1.1.12}$$

In other words, for all $x_r \leq x_{r-1} \leq \cdots \leq x_1$, with $x_r \leq 2$,

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{(m(n))^2} \log \mathbb{P} \Bigg(& \frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x_1, \cdots, \\ & \frac{R_r(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x_r \Bigg) = -K(x_r), \end{aligned} \tag{1.1.13}$$

while for all $2 \leq x_r \leq x_{r-1} \leq \cdots \leq x_1$,

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{(m(n))^2} \log \mathbb{P} \Bigg(& \frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x_1, \cdots, \\ & \frac{R_r(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x_r \Bigg) = 0. \end{aligned} \tag{1.1.14}$$

The LDP for the longest increasing subsequence is now a simple consequence:

Corollary 1.1.1 *Let m and n simultaneously converge to infinity in such a way that $m(n) = o(n^{1/4})$, then for any $x \geq 2$,*

$$\lim_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}} \geq x \right) = -2 \int_2^x \sqrt{(z/2)^2 - 1} dz,$$

and similarly, if $m(n) = o(n^{1/6})$, for any $x \leq 2$,

$$\lim_{n\to\infty} \frac{1}{(m(n))^2} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - n/m(n)}{\sqrt{n}} \leq x \right) = -K(x).$$

Remark 1.1.2 *The methodologies developed in this paper also allow to derive LDPs in related problems. Such is the case for last-passage directed percolation close to the x-axis, or for the departure time from many queues in series when the number of customers is a fractional power of the number of servers. In these two problems, similar discrete functional representations are available but with iid weights, so the large deviations rate functions should be the corresponding rate functions of the largest eigenvalue of the GUE.*

When the independent random letters are no longer uniformly drawn, let the $X_i^m, 1 \leq i \leq n$, be independently and identically distributed with $\mathbb{P}(X_1^m = \alpha_j) = p_j^m$, $1 \leq j \leq m$. Moreover, let $p_{max}^m = \max_{1\leq j\leq m} p_j^m$, let $p_{2nd}^m = \max\{p_j^m < p_{max}^m : 1 \leq j \leq m\}$, and let also $J(m) = \{j : p_j^m = p_{max}^m\}$, with $k(m) = card(J(m))$, *i.e.*, $k(m)$ is the multiplicity of p_{max}^m.

Theorem 1.1.3 *In the nonuniform case, let $k(m(n))$ and n simultaneously converge to infinity in such a way that $k(m(n))^3/p_{max}^m = o(n)$. Let also*

$$\frac{n(p_{2nd}^m)^2}{p_{max}^m} = o(\exp(-k(m(n))^\alpha)), \quad \text{for some } \alpha > 1, \tag{1.1.15}$$

then

$$\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}}$$

satisfies a LDP on $\mathbb{R}$ *with speed* $k(m(n))$ *and good rate function* I_1.

In other words, for any $x \geq 2$,

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \geq x \right) = -2\int_2^x \sqrt{(z/2)^2 - 1}dz, \quad (1.1.16)$$

while for any $x < 2$

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \leq x \right) = -\infty. \quad (1.1.17)$$

Above, the conditions on p_{max}^m match exactly those of Theorem 1.1.1.

When the renormalized variable is on the left of its simultaneous asymptotic mean, again we get a more accurate form of (1.1.17). Before presenting this result, let us first recall a few facts. For the alphabet $\mathcal{A}_m$ with corresponding probability set $\mathcal{P} = \{p_1^m, p_2^m, \cdots, p_m^m\}$, let $p^{(1)} > p^{(2)} > \cdots > p^{(l)}$, $1 \leq l \leq m$, be the distinct elements in $\mathcal{P}$, and let $d_1, \cdots, d_l$ be the corresponding multiplicities, with $\sum_{i=1}^l d_i = m$. Then $p^{(1)} = p_{max}^m$ and $d_1 = k(m)$ as in the previous notations. Let $\mathcal{G}_m(d_1, \cdots, d_l)$ be the set of $m \times m$ random matrices $\mathbf{X}$ which are direct sums of mutually independent elements of the $d_i \times d_i$ GUE, $1 \leq i \leq l$. Moreover, let $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)}$ be the non-increasing rearrangement of $\mathcal{P}$. The "generalized" $m \times m$ traceless GUE associated with $\mathcal{P}$ is the set, denoted by $\mathcal{G}^0(p_1^m, p_2^m, \cdots, p_m^m)$, of $m \times m$ random matrices $\mathbf{X^0}$, of the form

$$\mathbf{X}_{i,j}^0 = \begin{cases} \mathbf{X}_{i,i} - \sqrt{p_{(i)}} \sum_{h=1}^m \sqrt{p_{(h)}} \mathbf{X}_{h,h}, & \text{if } i = j, \\ \mathbf{X}_{i,j}, & \text{otherwise,} \end{cases} \quad (1.1.18)$$

where $\mathbf{X} \in \mathcal{G}_m(d_1, \cdots, d_l)$. Let $\tilde{\lambda}_1^0$ be the largest eigenvalue of the diagonal block corresponding to $p^{(1)} = p_{max}^m$ in $\mathbf{X^0}$.

Theorem 1.1.4 *Let* $k(m(n))$ *and* n *simultaneously converge to infinity in such a way that* $k(m(n))^5/p_{max}^m = o(n)$, *let*

$$\frac{n(p_{2nd}^m)^2}{p_{max}^m} = o(\exp(-k(m(n))^\alpha)), \qquad \textit{for some } \alpha > 2, \quad (1.1.19)$$

and assume that for some $0 \leq \eta \leq 1$,

$$\lim_{n\to\infty} k(m(n))p_{max}^m = \eta. \tag{1.1.20}$$

Then

$$\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}}$$

satisfies a LDP on $\mathbb{R}$ *with speed* $(k(m(n)))^2$ *and good rate function* K_η, *where* K_η *is the rate function of* $\tilde{\lambda}_1^0$ *when on the left of its asymptotic mean.*

In other words, for any $x \leq 2$,

$$\lim_{n\to\infty} \frac{1}{(k(m(n)))^2} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \leq x \right) = -K_\eta(x), \tag{1.1.21}$$

while for any $x \geq 2$,

$$\lim_{n\to\infty} \frac{1}{(k(m(n)))^2} \log \mathbb{P}\left(\frac{R_1(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \leq x \right) = 0. \tag{1.1.22}$$

Remark 1.1.3 *The rate function* K_η *is given by*

$$K_\eta(x) = \sup_{y\leq 0} \left(xy - yS(y) + J(S(y)) + \frac{\eta y^2}{2} \right),$$

where J *is the rate function (with speed* m^2*) of the largest eigenvalue of the* $m \times m$ *GUE, and for each* $y \leq 0$, $S(y)$ *is the unique solution to* $J'(t) = y$ *with* $t \leq 2$. *For* $x \geq 2$, $J(x) = 0$, *while for* $x \leq 2$, *the following closed form expression for* J *is obtained in [15],*

$$J(x) = \frac{1}{216}\Big(- x\left(-72x + x^3 + 30\sqrt{12 + x^2} + x^2\sqrt{12 + x^2}\right) \\ - 216 \log\left(\frac{1}{6}\left(x + \sqrt{12 + x^2}\right)\right)\Big). \tag{1.1.23}$$

In particular, $K_0 = J$ *and* $K_1 = K$. *In fact the relationship between the spectrum of GUE and traceless GUE implies that*

$$K(x) = \sup_{y\leq 0} \left(xy - J^*(y) + \frac{y^2}{2} \right),$$

where $*$ *denotes the Legendre transform. For any* $0 \leq \eta \leq 1$, $K_\eta(x) = 0$, *when* $x \geq 2$. *For* $0 \leq \eta < 1$ *and* $x \in (-\infty, 2)$, $K_\eta(x) > 0$ *and is asymptotically equivalent to*

$$\frac{x^2}{2(1-\eta)} + \log\left(-\frac{x}{1-\eta}\right),$$

as $x \to -\infty$. *For* $\eta = 1$, *when* $0 < x < 2$, $K_1(x) = K(x)$ *is positive and finite. As* $x \to 0$, $K(x) \sim -\log x$, *while as* $x \to 2$, $K(x) \sim C(2-x)^3$, *for some positive constant* C.

To complement the previous results, we provide corresponding concentration results. These rely in part on the concentration results for the largest eigenvalue of the $m \times m$ GUE matrix, obtained respectively in [3] and [27]. Comparing the forthcoming result with Corollary 1.1.1, we see that the deviation rates match the fluctuation results in this case. In turn these rates match the order of the tails of the Tracy-Widom distribution.

Theorem 1.1.5 *In the uniform model, let* $0 < \alpha < 1/4$, *and let* $m \leq An^\alpha$, *for some* $A > 0$. *Then for any* $0 < \epsilon < 1$,

$$\mathbb{P}\left(\frac{R_1(n,m) - n/m}{\sqrt{n/m}} \geq 2\sqrt{m}(1+\epsilon)\right) \leq C(A,\alpha)\exp\left\{-\frac{m\epsilon^{3/2}}{C(A,\alpha)}\right\}, \tag{1.1.24}$$

where

$$C(A,\alpha) = C\max\{A^{10/3}, 1\}\frac{1+\alpha}{1-4\alpha}\exp\left\{\frac{1+\alpha}{1-4\alpha}\right\},$$

for some absolute constant $C > 0$.

Likewise, let $0 < \alpha < 1/6$, *and let* $m \leq An^\alpha$, *for some* $A > 0$. *Then for any* $0 < \epsilon < 1$,

$$\mathbb{P}\left(\frac{R_1(n,m) - n/m}{\sqrt{n/m}} \leq 2\sqrt{m}(1-\epsilon)\right) \leq C(A,\alpha)\exp\left\{-\frac{m^2\epsilon^3}{C(A,\alpha)}\right\}, \tag{1.1.25}$$

where

$$C(A,\alpha) = C\max\{A^4, 1\}\frac{1+\alpha}{1-6\alpha}\exp\left\{\frac{1+\alpha}{1-6\alpha}\right\},$$

for some absolute constant $C > 0$.

Again, in the non-uniform case, we have similar results but under a further control of the second highest probability.

Theorem 1.1.6 *In the non-uniform model, let $\alpha > 3$, and let $k(m(n))^\alpha / p_{max}^m \leq An$, for some $A > 0$. Moreover, let*

$$\frac{n(p_{2nd}^m)^2}{p_{max}^m} \leq B \exp(-k(m(n))), \tag{1.1.26}$$

for some $B > 0$, then for any $0 < \epsilon < 1$,

$$\mathbb{P}\left(\frac{R_1(n,m) - np_{max}^m}{\sqrt{nk(m)p_{max}^m}} \geq 2(1+\epsilon) \right) \leq C(A,B,\alpha) \exp\left\{ -\frac{k(m)\epsilon^{3/2}}{C(A,B,\alpha)} \right\}, \tag{1.1.27}$$

where

$$C(A,B,\alpha) = C \max\{A^{10/3\alpha}, 1\} \max\{\sqrt{B}, 1\} \frac{\alpha+2}{\alpha-3} \exp\left\{ \frac{\alpha+2}{\alpha-3} \right\},$$

for some absolute constant $C > 0$.

Likewise, let $\alpha > 5$ and let $k(m(n))^\alpha / p_{max}^m \leq An$, with some $A > 0$, and let

$$\frac{n(p_{2nd}^m)^2}{p_{max}^m} \leq B \exp(-k(m(n))^2), \tag{1.1.28}$$

for some $B > 0$, then for any $0 < \epsilon < 1$,

$$\mathbb{P}\left(\frac{R_1(n,m) - np_{max}^m}{\sqrt{nk(m)p_{max}^m}} \leq 2(1-\epsilon) \right) \leq C(A,B,\alpha) \exp\left\{ -\frac{k(m)^2\epsilon^3}{C(A,B,\alpha)} \right\}, \tag{1.1.29}$$

where

$$C(A,B,\alpha) = C \max\{A^{4/\alpha}, 1\} \max\{\sqrt{B}, 1\} \frac{\alpha+2}{\alpha-5} \exp\left\{ \frac{\alpha+2}{\alpha-5} \right\},$$

for some absolute constant $C > 0$.

1.2 Proof of Theorem 1.1.1 and Theorem 1.1.2

As in [11], let

$$X_{i,j}^m = \begin{cases} 1, & \text{if } X_i^m = \alpha_j, \\ 0, & \text{otherwise,} \end{cases} \tag{1.2.1}$$

be Bernoulli random variables with parameter $1/m$. For a fixed $1 \leq j \leq m$, the $X_{i,j}^m$s are iid while for $j \neq j'$, $(X_{1,j}^m, \cdots, X_{n,j}^m)$ and $(X_{1,j'}^m, \cdots, X_{n,j'}^m)$ are identically distributed but no longer independent.

Let $S_k^{m,j} = \sum_{i=1}^k X_{i,j}^m$ be the number of occurences of α_j among $(X_i^m)_{1 \leq i \leq k}$. Since for $1 \leq k < l \leq n$, the number of occurences of α_j among $(X_i^m)_{k+1 \leq i \leq l}$ is $S_l^{m,j} - S_k^{m,j}$,

$$R_1(n,m) = \sup_{0=l_0 \leq l_1 \leq \cdots \leq l_m = n} \sum_{j=1}^m (S_{l_j}^{m,j} - S_{l_{j-1}}^{m,j}),$$

with the convention that $S_0^{m,j} = 0$.

Moreover, letting $V_k(n,m) = \sum_{i=1}^k R_i(n,m)$, combinatorial arguments yield (see Theorem 3.1 in [20])

$$V_k(n,m) = \sup_{\mathbf{t} \in I_{k,m}(n)} \sum_{j=1}^k \sum_{l=j}^{m-k+j} \left(S_{[t_{j,l}]}^{m,l} - S_{[t_{j,l-1}]}^{m,l} \right), \quad 1 \leq k \leq m, \tag{1.2.2}$$

where

$$\begin{aligned} I_{k,m}(n) = \{ \mathbf{t} = (t_{j,l} : 1 \leq j \leq k, 0 \leq l \leq m) : \\ t_{j,j-1} = 0, t_{j,m-k+j} = n, 1 \leq j \leq k; t_{j,l-1} \leq t_{j,l}, 1 \leq j \leq k, 1 \leq l \leq m-1; \\ t_{j,l} \leq t_{j-1,l-1}, 2 \leq j \leq k, 2 \leq l \leq m \}. \end{aligned}$$

Let $\tilde{X}_{i,j}^m = (X_{i,j}^m - 1/m)/\sigma_m$, with $\sigma_m^2 = (1/m)(1-1/m)$, let $\tilde{S}_k^{m,j} = \sum_{i=1}^k \tilde{X}_{i,j}^m$. Similarly define $\tilde{V}_k(n,m)$, $1 \leq k \leq m$ and let $\tilde{R}_k(n,m) = \tilde{V}_k(n,m) - \tilde{V}_{k-1}(n,m)$, $2 \leq k \leq m$, while $\tilde{R}_1(n,m) = \tilde{V}_1(n,m)$. Clearly $V_k(n,m) = \sigma_m \tilde{V}_k(n,m) + kn/m$, and

$$\frac{R_k(n,m) - n/m}{\sqrt{n}} = \sqrt{1 - \frac{1}{m}} \frac{\tilde{R}_k(n,m)}{\sqrt{nm}}.$$

Let

$$\tilde{V}_k(n,m) = \sup_{\mathbf{t} \in I_{k,m}(n)} \sum_{j=1}^k \sum_{l=j}^{m-k+j} \left(\tilde{S}_{[t_{j,l}]}^{m,l} - \tilde{S}_{[t_{j,l-1}]}^{m,l} \right), \quad 1 \leq k \leq m, \tag{1.2.3}$$

with

$$Cov(\tilde{S}_k^{m,i}, \tilde{S}_k^{m,j}) = \begin{cases} k, & \text{if } i = j, \\ k\rho, & \text{otherwise,} \end{cases} \tag{1.2.4}$$

and $\rho = -1/(m-1)$.

Next, $\tilde{V}_k(n,m)$ can be approximated by

$$\tilde{L}_k(n,m) = \sup_{\mathbf{t}\in I_{k,m}(n)} \sum_{j=1}^{k} \sum_{l=j}^{m-k+j} \left(\tilde{B}^l_{t_{j,l}} - \tilde{B}^l_{t_{j,l-1}} \right), \quad 1 \le k \le m, \tag{1.2.5}$$

where $(\tilde{B}^j)_{1\le j\le m}$ is a driftless m-dimensional Brownian Motion with covariance matrix given in (1.1.2), and

$$\tilde{L}_k(n,m) \stackrel{\mathcal{L}}{=} \sqrt{n}\tilde{L}_k(1,m).$$

More precisely, inspired by [12],

$$|\tilde{V}_k(n,m) - \tilde{L}_k(n,m)| \le 2k \sum_{l=1}^{m} (Y_n^{m,l} + W_n^l), \tag{1.2.6}$$

where

$$Y_n^{m,l} = \max_{1\le i\le n} |\tilde{S}_i^{m,l} - \tilde{B}_i^l| \quad \text{and} \quad W_n^l = \sup_{\substack{0\le s,t\le n \\ |s-t|\le 1}} |\tilde{B}_s^l - \tilde{B}_t^l|.$$

Since

$$\left(\tilde{R}_k(n,m)\right)_{1\le i\le m} = \mathbf{\Theta}_m^{-1} \left(\left(\tilde{V}_k(n,m)\right)_{1\le k\le m} \right),$$

for any $\epsilon > 0$, and from (1.2.6),

$$\begin{aligned}
&\mathbb{P}\left(\left| \tilde{R}_k(n,m) - \left(\tilde{L}_k(n,m) - \tilde{L}_{k-1}(n,m) \right) \right| \ge \sqrt{mn}\epsilon \right) \\
&\le \mathbb{P}\left(2(2k-1) \sum_{l=1}^{m} (Y_n^{m,l} + W_n^l) \ge \sqrt{mn}\epsilon \right) \\
&\le \mathbb{P}\left(\sum_{l=1}^{m} Y_n^{m,l} \ge \frac{\sqrt{mn}\epsilon}{4(2k-1)} \right) + \mathbb{P}\left(\sum_{l=1}^{m} W_n^l \ge \frac{\sqrt{mn}\epsilon}{4(2k-1)} \right) \\
&\le \sum_{l=1}^{m} \left(\mathbb{P}\left(Y_n^{m,l} \ge \frac{\sqrt{mn}\epsilon}{m(8k-4)} \right) + \mathbb{P}\left(W_n^l \ge \frac{\sqrt{mn}\epsilon}{m(8k-4)} \right) \right) \\
&= m\mathbb{P}\left(Y_n^{m,1} \ge \frac{\sqrt{n}\epsilon}{\sqrt{m}(8k-4)} \right) + m\mathbb{P}\left(W_n^1 \ge \frac{\sqrt{n}\epsilon}{\sqrt{m}(8k-4)} \right),
\end{aligned} \tag{1.2.7}$$

for $1 \le k \le m$, and with the convention that $\tilde{L}_0(n,m) = 0$.

From Sakhanenko's version of the KMT inequality as stated, for example, in Theorem 2.1 and Corollary 3.2 of [28],

$$\mathbb{P}\left(Y_n^{m,1} \ge \frac{\sqrt{n}\epsilon}{\sqrt{m}(8k-4)} \right) \le (1 + c_2(m)\sqrt{n}) \exp\left\{ -c_1(m) \frac{\sqrt{n}\epsilon}{\sqrt{m}(8k-4)} \right\}, \tag{1.2.8}$$

where, as $m \to +\infty$, $c_1(m) \sim C_1/\sqrt{m}$ and $c_2(m) \sim C_2/\sqrt{m}$, for absolute constants C_1 and C_2. Moreover,

$$\begin{aligned}\mathbb{P}\left(W_n^1 \geq \frac{\sqrt{n}\epsilon}{\sqrt{m}(8k-4)}\right) &\leq 2n\mathbb{P}\left(|\tilde{B}_2^1| \geq \frac{\sqrt{n}\epsilon}{\sqrt{m}(16k-8)}\right) \\ &= 4n\mathbb{P}\left(\tilde{B}_2^1 \geq \frac{\sqrt{n}\epsilon}{\sqrt{m}(16k-8)}\right) \\ &\leq 4en\exp\left\{-\frac{n\epsilon^2}{4em(16k-8)^2}\right\}. \end{aligned} \tag{1.2.9}$$

Combining (1.2.8) and (1.2.9), under the condition $m(n) = o(n^{1/4})$,

$$\mathbb{P}\left(\left|\tilde{R}_k(n,m) - \left(\tilde{L}_k(n,m) - \tilde{L}_{k-1}(n,m)\right)\right| \geq \sqrt{mn}\epsilon\right) \leq C_3\sqrt{mn}\exp\left\{-\frac{\sqrt{n}\epsilon}{C_3 m}\right\}, \tag{1.2.10}$$

for $1 \leq k \leq r$, and where C_3 is a positive constant depending on k, which for r fixed, can be chosen only depending on r.

For any $x_1 \geq x_2 \cdots \geq x_r > 2$, $r \geq 1$, and $0 < \epsilon < (x_r - 2)$,

$$\begin{aligned}&\mathbb{P}\left(\frac{\tilde{R}_1(n,m)}{\sqrt{mn}} \geq x_1, \frac{\tilde{R}_2(n,m)}{\sqrt{mn}} \geq x_2, \cdots, \frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \geq x_r\right) \\ &\leq \mathbb{P}\left(\frac{\tilde{L}_1(n,m) - \tilde{L}_0(n,m)}{\sqrt{mn}} \geq x_1 - \epsilon, \cdots, \frac{\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m)}{\sqrt{mn}} \geq x_r - \epsilon\right) \\ &\qquad + \sum_{i=1}^{r}\mathbb{P}\left(\frac{\tilde{R}_i(n,m) - (\tilde{L}_i(n,m) - \tilde{L}_{i-1}(n,m))}{\sqrt{mn}} \geq \epsilon\right), \end{aligned} \tag{1.2.11}$$

and

$$\begin{aligned}&\mathbb{P}\left(\frac{\tilde{R}_1(n,m)}{\sqrt{mn}} \geq x_1, \frac{\tilde{R}_2(n,m)}{\sqrt{mn}} \geq x_2, \cdots, \frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \geq x_r\right) \\ &\geq \mathbb{P}\left(\frac{\tilde{L}_1(n,m) - \tilde{L}_0(n,m)}{\sqrt{mn}} \geq x_1 + \epsilon, \cdots, \frac{\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m)}{\sqrt{mn}} \geq x_r + \epsilon\right) \\ &\qquad - \sum_{i=1}^{r}\mathbb{P}\left(\frac{(\tilde{L}_i(n,m) - \tilde{L}_{i-1}(n,m)) - \tilde{R}_i(n,m)}{\sqrt{mn}} \geq \epsilon\right), \end{aligned} \tag{1.2.12}$$

with again the convention that $\tilde{L}_0(n,m) = 0$.

Combining (1.1.1) with Theorem 1.5.2 of the Appendix, when m and n simultaneously converge to infinity in such a way that $m(n) = o(n^{1/4})$, the large deviations

for $(\tilde{L}_k(n,m))_{1\le k\le r}$ are then given by:

$$\lim_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}}\ge x_1,\cdots,\frac{\tilde{L}_r(n,m(n))-\tilde{L}_{r-1}(n,m(n))}{\sqrt{m(n)n}}\ge x_r\right)$$
$$=-2\sum_{i=1}^{r}\int_2^{x_i}\sqrt{(z/2)^2-1}dz, \quad (1.2.13)$$

for all $x_1\ge x_2\ge\cdots\ge x_r>2$. This implies that,

$$\mathbb{P}\left(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}}\ge x_1\pm\epsilon,\cdots,\frac{\tilde{L}_r(n,m(n))-\tilde{L}_{r-1}(n,m(n))}{\sqrt{m(n)n}}\ge x_r\pm\epsilon\right)$$
$$=\exp\left\{-m(n)\left(I_r(x_1\pm\epsilon,\cdots,x_r\pm\epsilon)+o(1)\right)\right\},$$

where $o(1)$ goes to 0 as n converges to infinity. Combining this fact with (1.2.10), for any $1\le k\le r$

$$\frac{\mathbb{P}\left(\left|\tilde{R}_k(n,m)-\left(\tilde{L}_k(n,m)-\tilde{L}_{k-1}(n,m)\right)\right|\ge\sqrt{mn}\epsilon\right)}{\mathbb{P}\left(\tilde{L}_1(n,m)\ge\sqrt{mn}(x_1\pm\epsilon),\cdots,\tilde{L}_r(n,m)-\tilde{L}_{r-1}(n,m)\ge\sqrt{mn}(x_r\pm\epsilon)\right)}$$
$$\le C_3\sqrt{mn}\exp\left\{-\frac{\sqrt{n}\epsilon}{C_3m}+m(I_r(x_1\pm\epsilon,\cdots,x_r\pm\epsilon)+o(1))\right\}$$
$$=C_3\sqrt{mn}\exp\left\{\frac{\sqrt{n}}{m}\left(-\frac{\epsilon}{C_3}+\frac{m^2}{\sqrt{n}}\left(I_r(x_1\pm\epsilon,\cdots,x_r\pm\epsilon)+o(1)\right)\right)\right\}$$
$$\to 0 \qquad \text{as } m,n\to\infty, \quad m=o(n^{1/4}). \quad (1.2.14)$$

From (1.2.11) and (1.2.14), as m and n simultaneously converge to infinity with $m=o(n^{1/4})$,

$$\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}}\ge x_1,\cdots,\frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}}\ge x_r\right) \quad (1.2.15)$$
$$\le\limsup_{n\to\infty}\frac{1}{m(n)}\log 2\mathbb{P}\left(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}}\ge x_1-\epsilon,\cdots,\right.$$
$$\left.\frac{\tilde{L}_r(n,m(n))-\tilde{L}_r(n,m(n))}{\sqrt{m(n)n}}\ge x_r-\epsilon\right)$$
$$=-I_r(x_1-\epsilon,\cdots,x_r-\epsilon).$$

Likewise, from (1.2.12) and (1.2.14),

$$\liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \geq x_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \geq x_r \right) \tag{1.2.16}$$

$$\geq \liminf_{n\to\infty} \frac{1}{m(n)} \log \frac{1}{2} \mathbb{P}\Bigg(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}} \geq x_1 + \epsilon, \cdots, \frac{\tilde{L}_r(n,m(n)) - \tilde{L}_r(n,m(n))}{\sqrt{m(n)n}} \geq x_r + \epsilon \Bigg)$$

$$= -I_r(x_1 + \epsilon, \cdots, x_r + \epsilon).$$

Now letting $\epsilon \to 0$,

$$\lim_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \geq x_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \geq x_r \right)$$
$$= -2 \sum_{i=1}^{r} \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz,$$

for any $x_1 \geq x_2 \cdots \geq x_r > 2$. Next, assume that $x_1 \geq x_2 \cdots \geq x_k > x_{k+1} = \cdots = x_r = 2$, $1 \leq k \leq r$, with the convention that $k = r$ corresponds to $x_1 \geq x_2 \cdots \geq x_r > 2$. Under the conditions given in Theorem 1.1.1, for any $\epsilon > 0$

$$\liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \geq x_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \geq x_r \right)$$
$$\geq -2 \sum_{i=1}^{k} \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz - 2 \sum_{i=k+1}^{r} \int_2^{2+\epsilon} \sqrt{(z/2)^2 - 1} dz.$$

Letting $\epsilon \to 0$, gives

$$\liminf_{n\to\infty} \frac{1}{m(n)} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \geq x_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \geq x_r \right)$$
$$\geq -2 \sum_{i=1}^{r} \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz, \tag{1.2.17}$$

while,

$$\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}}\geq x_1,\cdots,\frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}}\geq x_r\right)$$
$$\leq\limsup_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}}\geq x_1,\cdots,\frac{\tilde{R}_k(n,m(n))}{\sqrt{m(n)n}}\geq x_k\right)$$
$$=-2\sum_{i=1}^{r}\int_2^{x_i}\sqrt{(z/2)^2-1}dz. \tag{1.2.18}$$

Combining (1.2.17) and (1.2.18), (1.1.6) is proved.

Now fix $x<2$, let $0<\epsilon<(2-x)$, then

$$\mathbb{P}\left(\frac{\tilde{R}_k(n,m)}{\sqrt{mn}}\leq x\right)\leq\mathbb{P}\left(\frac{\tilde{L}_k(n,m)-\tilde{L}_{k-1}(n,m)}{\sqrt{mn}}\leq x+\epsilon\right)$$
$$+\mathbb{P}\left(\frac{|\tilde{R}_k(n,m)-(\tilde{L}_k(n,m)-\tilde{L}_{k-1}(n,m))|}{\sqrt{mn}}\geq\epsilon\right), \tag{1.2.19}$$

for any $1\leq k\leq r$. From (1.2.24), the first term on the right of (1.2.19) is exponentially negligible with speed m. For the second term, from (1.2.10), for any $T>0$,

$$\mathbb{P}\left(\left|\tilde{R}_k(n,m)-\left(\tilde{L}_k(n,m)-\tilde{L}_{k-1}(n,m)\right)\right|\geq\sqrt{mn}\epsilon\right)e^{mT} \tag{1.2.20}$$
$$\leq C_3\sqrt{mn}\exp\left\{\frac{\sqrt{n}}{m}\left(-\frac{\epsilon}{C_3}+\frac{m^2T}{\sqrt{n}}\right)\right\}\to 0\qquad\text{as } m,n\to\infty,\quad m=o(n^{1/4}).$$

Next, letting $T\to\infty$, we obtain that, for any $x<2$ and $1\leq k\leq r$,

$$\lim_{n\to\infty}\frac{1}{m(n)}\log\mathbb{P}\left(\frac{\tilde{R}_k(n,m(n))}{\sqrt{m(n)n}}\leq x\right)=-\infty, \tag{1.2.21}$$

which proves (1.1.7) in Theorem 1.1.1. ∎

Proof of Theorem 1.1.2

First, (1.1.14) is just a direct consequence of (1.1.6). Next, we prove (1.1.13). Fix $y_1\geq y_2\geq\cdots\geq y_r$, with $y_r<2$. If $K(y_r)<+\infty$, then there exists $\delta>0$ such that

$K(y_r - \delta) < +\infty$ and such that for any $0 < \epsilon < \min\{\delta, 2 - y_r\}$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\tilde{R}_1(n,m)}{\sqrt{mn}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \leq y_r\right) \qquad (1.2.22)\\
&\leq \mathbb{P}\left(\frac{\tilde{L}_1(n,m) - \tilde{L}_0(n,m)}{\sqrt{mn}} \leq y_1 + \epsilon, \cdots, \frac{\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m)}{\sqrt{mn}} \leq y_r + \epsilon\right)\\
&\qquad + \sum_{i=1}^{r} \mathbb{P}\left(\frac{|\tilde{R}_i(n,m) - (\tilde{L}_i(n,m) - \tilde{L}_{i-1}(n,m))|}{\sqrt{mn}} \geq \epsilon\right),
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\frac{\tilde{R}_1(n,m)}{\sqrt{mn}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \leq y_r\right) \qquad (1.2.23)\\
&\geq \mathbb{P}\left(\frac{\tilde{L}_1(n,m) - \tilde{L}_0(n,m)}{\sqrt{mn}} \leq y_1 - \epsilon, \cdots, \frac{\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m)}{\sqrt{mn}} \leq y_r - \epsilon\right)\\
&\qquad - \sum_{i=1}^{r} \mathbb{P}\left(\frac{|\tilde{R}_i(n,m) - (\tilde{L}_i(n,m) - \tilde{L}_{i-1}(n,m))|}{\sqrt{mn}} \geq \epsilon\right),
\end{aligned}$$

with once more the convention that $\tilde{L}_0(n,m) = 0$.

Combining (1.1.1) with Corollary 1.5.1, when m and n simultaneously converge to infinity with $m = o(n^{1/6})$,

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{L}_r(n,m(n)) - \tilde{L}_{r-1}(n,m(n))}{\sqrt{m(n)n}} \leq y_r\right)\\
&\qquad = -K(y_r), \quad (1.2.24)
\end{aligned}$$

for all $y_r \leq y_{r-1} \leq \cdots \leq y_1$ with $y_r < 2$. Thus

$$\begin{aligned}
&\mathbb{P}\left(\frac{\tilde{L}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1 \pm \epsilon, \cdots, \frac{\tilde{L}_r(n,m(n)) - \tilde{L}_{r-1}(n,m(n))}{\sqrt{m(n)n}} \leq y_r \pm \epsilon\right)\\
&\qquad = \exp\left\{-m(n)^2 \left(K(y_r \pm \epsilon) + o(1)\right)\right\},
\end{aligned}$$

where $o(1)$ is meant for an expression converging to zero as n converges to infinity.

Combining this last fact with (1.2.10), for any $1 \leq k \leq r$

$$\begin{aligned}
&\frac{\mathbb{P}\left(\left|\tilde{R}_k(n,m) - \left(\tilde{L}_k(n,m) - \tilde{L}_{k-1}(n,m)\right)\right| \geq \sqrt{mn}\epsilon\right)}{\mathbb{P}\left(\tilde{L}_1(n,m) \leq \sqrt{mn}(y_1 \pm \epsilon), \cdots, \tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m) \leq \sqrt{mn}(y_r \pm \epsilon)\right)} \qquad (1.2.25)\\
&\leq C_3\sqrt{mn}\exp\left\{\frac{\sqrt{n}}{m}\left(-\frac{\epsilon}{C_3} + \frac{m^3}{\sqrt{n}}\left(K(y_r \pm \epsilon) + o(1)\right)\right)\right\}\\
&\to 0 \qquad \text{as } m,n \to \infty, \quad m = o(n^{1/6}).
\end{aligned}$$

Repeating previous arguments, letting ϵ go to 0, and since $m = o(n^{1/6})$,

$$\lim_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \leq y_r\right) = -K(y_r), \quad (1.2.26)$$

for $y_r \leq y_{r-1} \leq \cdots \leq y_1$, with $y_r < 2$ and $K(y_r) < +\infty$.

Now for fixed $y_1 \geq y_2 \geq \cdots \geq y_r$, $y_r < 2$, we tackle the case $K(y_r) = +\infty$. Since

$$\begin{aligned}
&\mathbb{P}\left(\frac{\tilde{R}_1(n,m)}{\sqrt{mn}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \leq y_r\right) \leq \mathbb{P}\left(\frac{\tilde{R}_r(n,m)}{\sqrt{mn}} \leq y_r\right)\\
&\leq \mathbb{P}\left(\frac{\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m)}{\sqrt{mn}} \leq y_r + \epsilon\right) +\\
&\qquad\qquad \mathbb{P}\left(\frac{|\tilde{R}_r(n,m) - (\tilde{L}_r(n,m) - \tilde{L}_{r-1}(n,m))|}{\sqrt{mn}} \geq \epsilon\right), \qquad (1.2.27)
\end{aligned}$$

and when m and n simultaneously converge to infinity with $m = o(n^{1/6})$, the second term on the right of (1.2.27) is exponentially negligible with speed m^2, while the first term is, from (1.2.24), dominated by $e^{-m(n)^2 K(y_r+\epsilon)}$. Thus (1.2.26), in this case, follows by letting ϵ go to 0.

Now let $2 = y_r \leq y_{r-1} \leq \cdots \leq y_1$, then for any $\epsilon > 0$,

$$\begin{aligned}
&\liminf_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \leq y_r\right)\\
&\geq \liminf_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \leq 2 - \epsilon\right)\\
&= -K(2-\epsilon). \qquad (1.2.28)
\end{aligned}$$

Again, letting ϵ goes to zero, and since K is continuous (see the Appendix for a proof), we get

$$\liminf_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \leq y_r \right) \geq -K(2) = 0.$$

Clearly,

$$\limsup_{n\to\infty} \frac{1}{m(n)^2} \log \mathbb{P}\left(\frac{\tilde{R}_1(n,m(n))}{\sqrt{m(n)n}} \leq y_1, \cdots, \frac{\tilde{R}_r(n,m(n))}{\sqrt{m(n)n}} \leq y_r \right) \leq 0,$$

which proves the case $y_r = 2$, and finishes the proof of the first part of Theorem 1.1.2. From Lemma 1.5.2 of the Appendix, we can prove (1.1.11).

When $x \leq 0$, $\mathcal{M}_0((-\infty, x])$ is empty so $K(x) = +\infty$ and when $x \geq 2$, the semicircular probability measure belongs to $\mathcal{M}_0((-\infty, x])$, thus $K(x) = 0$. When $0 < x < 2$, the closed form expression of K given by (1.1.12) can be derived using the techniques developed in [33]. Denote by μ_0 the zero mean probability measure supported on $(-\infty, x]$, minimizing

$$I(\mu) = \frac{1}{2} \int y^2 d\mu(y) - \iint \log|t-y| d\mu(t) d\mu(y) - \frac{3}{4}, \tag{1.2.29}$$

(the existence and uniqueness of μ_0 follows from Theorem 1.3 in Chapter 1 of [35]. Moreover, in view of Theorem 2.5 in Chapter IV of [35], μ_0 is absolutely continuous with continuous density ρ_0, while from Theorem 1.10 and Theorem 1.11 of Chapter IV in [35], its support is a finite interval). Let us now proceed to explicitly find ρ_0. To do so, consider the Lagrange function

$$E(\mu) = I(\mu) + c_1 \left(\int d\mu(y) - 1 \right) + c_2 \int y d\mu(y),$$

where the Lagrange multipliers c_1 and c_2 correspond to the constraints that μ is a zero mean probability measure. Let $[L', x]$ be the support of ρ_0, and for any continuous

function h supported on $[L', x]$ such that $h(y) \geq -\rho_0(y)$ let

$$\begin{aligned} E(\rho_0 + \epsilon h) = \frac{1}{2} \int y^2 \left(\rho_0(y) + \epsilon h(y)\right) dy \\ - \iint \log|t-y| \left(\rho_0(t) + \epsilon h(t)\right) \left(\rho_0(y) + \epsilon h(y)\right) dtdy - \frac{3}{4} \\ + c_1 \left(\int \left(\rho_0(y) + \epsilon h(y)\right) dy - 1 \right) + c_2 \int y \left(\rho_0(y) + \epsilon h(y)\right) dy. \end{aligned} \tag{1.2.30}$$

Thus

$$\left. \frac{dE(\rho_0 + \epsilon h)}{d\epsilon} \right|_{\epsilon=0} = 0,$$

gives

$$\int \left(\frac{y^2}{2} - 2 \int \log|t-y| \rho_0(t) dt + c_1 + c_2 y \right) h(y) dy = 0, \tag{1.2.31}$$

for any continuous h such that $h(y) \geq -\rho_0(y)$. Let

$$g(y) = \frac{y^2}{2} - 2 \int \log|t-y| \rho_0(t) dt + c_1 + c_2 y,$$

which is a continuous function on $[L', x]$. Let $h(y) = g^+(y)$, then (1.2.31) yields that

$$\int_{g(y) \geq 0} g(y)^2 dy = 0,$$

thus $g(y) \leq 0$ for $y \in [L', x]$. Likewise, letting

$$h(y) = \begin{cases} 0, & \text{if } g(y) > 0, \\ g(y), & \text{if } -\rho_0(y) \leq g(y) \leq 0, \\ -\rho_0(y), & \text{if } g(y) < -\rho_0(y), \end{cases} \tag{1.2.32}$$

then (1.2.31) yields that $g(y) \geq 0$ for $y \in [L', x]$. Thus

$$\frac{y^2}{2} - 2 \int \log|t-y| \rho_0(t) dt + c_1 + c_2 y = 0, \tag{1.2.33}$$

for any $y \in [L', x]$. In turn, differentiating (1.2.33) with respect to y further gives,

$$y - 2 \text{ p.v.} \int \frac{\rho_0(t)}{y-t} dt + c_2 = 0, \tag{1.2.34}$$

where p.v. is the Cauchy principal value.

Let $L = L' - x$ and $f_x(t) = \rho_0(t+x)$ be supported on $[L, 0]$, then the finite Hilbert transform

$$\frac{1}{\pi}\text{p.v.}\int_{L'}^{x}\frac{\rho_0(t)}{y-t}dt = \frac{y+c_2}{2\pi},$$

becomes

$$\frac{1}{\pi}\text{p.v.}\int_{L}^{0}\frac{f_x(t)}{y-t}dt = \frac{x+y+c_2}{2\pi},$$

for any $y \in [L, 0]$. From Section 4.3 of [41], this finite Hilbert transform can be inverted as

$$f_x(y) = \frac{1}{\pi\sqrt{(y-L)(-y)}}\left(\text{p.v.}\int_{L}^{0}\frac{\sqrt{(t-L)(-t)}}{t-y}\frac{x+t+c_2}{2\pi}dt + c_3\right), \tag{1.2.35}$$

where $L \le y \le 0$. Moreover,

$$\begin{aligned}\text{p.v.}\int_{L}^{0}\frac{\sqrt{(t-L)(-t)}}{t-y}\frac{x+t+c_2}{2\pi}dt \\ = \frac{1}{16}\left(4c_2(L-2y) + L^2 + 4L(x+y) - 8y(x+y)\right).\end{aligned} \tag{1.2.36}$$

From $f_x(L) = 0$, we get

$$c_3 = \frac{1}{16}\left(4L(c_2+x) + 3L^2\right),$$

and plugging this into (1.2.35) yields

$$f_x(y) = \frac{\sqrt{y(L-y)}(2c_2 + L + 2(x+y))}{4\pi y}.$$

Now from the two constraints $\int d\mu_0(y) = 1$ and $\int y d\mu_0(y) = 0$ we get

$$\int_{L}^{0} y f_x(y)dy + x = 0, \int_{L}^{0} f_x(y)dy = 1,$$

which further gives

$$L = \frac{2\ 2^{2/3}\left(\sqrt{81x^2+12} - 9x\right)^{2/3} - 4\ 6^{1/3}}{3^{2/3}\left(\sqrt{81x^2+12} - 9x\right)^{1/3}}, \tag{1.2.37}$$

and

$$c_2 = \frac{2\ 3^{2/3} - \sqrt[3]{6}\left(\sqrt{81x^2+12}-9x\right)^{2/3}}{2^{2/3}\left(\sqrt{81x^2+12}-9x\right)^{1/3}} - \frac{\sqrt[3]{2}3^{2/3}\left(\sqrt{81x^2+12}-9x\right)^{2/3} + \frac{6\ 2^{2/3}\sqrt[3]{3}}{\left(\sqrt{81x^2+12}-9x\right)^{2/3}} + 6}{18x} - x. \quad (1.2.38)$$

Integrating (1.2.33) with respect to μ_0 gives,

$$\iint \log|y-t|d\mu_0(t)d\mu_0(y) = \frac{1}{4}\int y^2 d\mu_0(y) + \frac{c_1}{2},$$

while c_1 can be determined by substituting $y = x$ in (1.2.33),

$$c_1 = -\frac{x^2}{2} + 2\int \log|x-t|d\mu_0(t) - c_2 x.$$

Finally,

$$\begin{aligned} I(\mu_0) &= \frac{1}{2}\int y^2 d\mu_0(y) - \iint \log|t-y|d\mu_0(t)d\mu_0(y) - \frac{3}{4} \\ &= \frac{1}{4}\int_L^0 (x+y)^2 f_x(y)dy - \int_L^0 \log(-y) f_x(y)dy + \frac{x^2}{4} + \frac{c_2 x}{2} - \frac{3}{4}. \end{aligned} \quad (1.2.39)$$

Plugging L and c_2 into (1.2.39) gives the closed form expression for K. ∎

1.3 Proof of Theorem 1.1.3 and Theorem 1.1.4

Recall that

$$R_1(n,m) = V_1(n,m) = \sup_{0=l_0\le l_1\le\cdots\le l_m=n} \sum_{j=1}^{m} (S_{l_j}^{m,j} - S_{l_{j-1}}^{m,j}).$$

Then, let

$$V_1^{'}(n,m) = \sup_{\substack{0=l_0\le l_1\le\cdots\le l_m=n \\ l_{j-1}=l_j \text{ for } j\notin J(m)}} \sum_{j=1}^{m} (S_{l_j}^{m,j} - S_{l_{j-1}}^{m,j}),$$

where from Lemma 9 of [11],

$$\mathbb{E}\left(\left|V_1(n,m) - V_1^{'}(n,m)\right|\right) \le Cnp_{2nd}^m, \quad (1.3.1)$$

with $C > 0$ some absolute constant.

To prove Theorem 1.1.3, let us first prove a lemma,

Lemma 1.3.1 *Let $k(m(n))$ converge to infinity with n in such a way that $k(m(n))^3/p_{max}^m = o(n)$, then for any $x \geq 2$,*

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{V_1'(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \geq x \right) = -2\int_2^x \sqrt{(z/2)^2 - 1}dz, \quad (1.3.2)$$

and for any $x < 2$,

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{V_1'(n, m(n)) - np_{max}^m}{\sqrt{nk(m(n))p_{max}^m}} \leq x \right) = -\infty. \quad (1.3.3)$$

Proof.

As in the proof of Theorem 1.1.1, for any $j \in J(m)$, set $\tilde{X}_{i,j}^m = (X_{i,j}^m - p_{max}^m)/\sigma_m$, where $\sigma_m^2 = p_{max}^m(1 - p_{max}^m)$, and set $\tilde{S}_k^{m,j} = \sum_{i=1}^k \tilde{X}_{i,j}^m$. Hence

$$\frac{V_1'(n, m) - np_{max}^m}{\sqrt{nkp_{max}^m}} = (\sqrt{1 - p_{max}^m})\frac{\tilde{V}_1'(n, m)}{\sqrt{nk}},$$

with the obvious notation for $\tilde{V}_1'(n, m)$. Since $k(m(n))p_{max}^m \leq 1$, as $n \to \infty$, $p_{max}^m \to 0$, so (1.3.2) can be reduced to,

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{\tilde{V}_1'(n, m(n))}{\sqrt{nk(m(n))}} \geq x \right) = -I_1(x), \quad (1.3.4)$$

for any $x \geq 2$.

Moreover, (1.3.3) can be reduced to,

$$\lim_{n\to\infty} \frac{1}{k(m(n))} \log \mathbb{P}\left(\frac{\tilde{V}_1'(n, m(n))}{\sqrt{nk(m(n))}} \leq x \right) = -\infty, \quad (1.3.5)$$

for any $x < 2$.

Since

$$\tilde{V}_1'(n, m) = \sup_{\substack{0=l_0\leq l_1\leq\cdots\leq l_m=n \\ l_{j-1}=l_j \text{ for } j\notin J(m)}} \sum_{j=1}^m (\tilde{S}_{l_j}^{m,j} - \tilde{S}_{l_{j-1}}^{m,j}), \quad (1.3.6)$$

with

$$Cov(\tilde{S}_k^{m,i}, \tilde{S}_k^{m,j}) = \begin{cases} k, & \text{if } i = j, \\ k\rho_1, & \text{otherwise,} \end{cases} \quad (1.3.7)$$

where $\rho_1 = -p_{max}^m/(1-p_{max}^m)$, it can be approximated via KMT by the Brownian functional $F(n,k)$

$$F(n,k) = \sup_{0=t_0\le t_1\le\cdots\le t_k=n} \sum_{r=1}^{k} (\tilde{B}_{t_r}^{(r)} - \tilde{B}_{t_{r-1}}^{(r)}), \tag{1.3.8}$$

where $(\tilde{B}^{(r)})_{1\le r\le k}$ is a k-dimensional Brownian motion with covariance matrix

$$t \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \cdots & 1 \end{pmatrix}.$$

Moreover,

$$F(n,k) \overset{\mathcal{L}}{=} \sqrt{n} F(1,k), \tag{1.3.9}$$

while from Corollary 3.2 and Corollary 3.3 in [19],

$$\begin{aligned} &\sqrt{1-p_{max}^m} F(1,k) \overset{\mathcal{L}}{=} \\ &\frac{\sqrt{1-kp_{max}^m}-1}{k} \sum_{j=1}^{k} B_1^j + \sup_{0=t_0\le t_1\le\cdots\le t_k=1} \sum_{r=1}^{k} (B_{t_r}^r - B_{t_{r-1}}^r), \end{aligned} \tag{1.3.10}$$

where $(B^j)_{1\le j\le k}$ is a standard k-dimensional Brownian motion. Looking at the right hand side of (1.3.10), the first sum is a Gaussian random variable with variance at most $1/k$, while for the second part, it is well known that:

$$\sup_{0=t_0\le t_1\le\cdots\le t_k=1} \sum_{r=1}^{k} (B_{t_r}^r - B_{t_{r-1}}^r) \overset{\mathcal{L}}{=} \lambda_1^k, \tag{1.3.11}$$

where λ_1^k is the largest eigenvalue of a $k\times k$ element of the GUE (see the Introduction). For the large deviation of $F(1,k)$ when it is on the left of its asymptotic mean, since $\lambda_1^k/\sqrt{k}$ satisfies a LDP with rate function I_1 and since the contribution of the Gaussian term is negligible, we get, as shown in the Appendix, that

$$\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}(F(1,k) \ge \sqrt{k}x) = -I_1(x). \tag{1.3.12}$$

Next, as in the proof of Theorem 1.1.1,

$$\mathbb{P}\left(\left|\tilde{V}_1'(n,m) - F(n,k)\right| \geq \sqrt{nk}\epsilon\right) \leq k\left(\mathbb{P}\left(Y_n^{m,l} \geq \frac{\sqrt{n}\epsilon}{4\sqrt{k}}\right) + \mathbb{P}\left(W_n^l \geq \frac{\sqrt{n}\epsilon}{4\sqrt{k}}\right)\right), \quad (1.3.13)$$

where l is any element of $J(m)$ and

$$Y_n^{m,l} = \max_{1\leq i\leq n} |\tilde{S}_i^{m,l} - \tilde{B}_i^{(l)}| \quad \text{and} \quad W_n^l = \sup_{\substack{0\leq s,t\leq n \\ |s-t|\leq 1}} |\tilde{B}_s^{(l)} - \tilde{B}_t^{(l)}|.$$

As in getting (1.2.8), we have

$$\mathbb{P}\left(Y_n^{m,1} \geq \frac{\sqrt{n}\epsilon}{4\sqrt{k}}\right) \leq (1 + c_2(p_{max}^m)\sqrt{n}) \exp\left\{-c_1(p_{max}^m)\frac{\sqrt{n}\epsilon}{4\sqrt{k}}\right\}, \quad (1.3.14)$$

where $c_1(p_{max}^m) \sim C_1\sqrt{p_{max}^m}$ and $c_2(m) \sim C_2\sqrt{p_{max}^m}$, for some constants C_1 and C_2, and from (1.2.9),

$$\mathbb{P}\left(W_n^1 \geq \frac{\sqrt{n}\epsilon}{4\sqrt{k}}\right) \leq C_3 n \exp\left\{-\frac{n\epsilon^2}{C_3 k}\right\}, \quad (1.3.15)$$

for some positive constant C_3. Combining (1.3.14) and (1.3.15), under the condition $k(m(n))^3/p_{max}^m = o(n)$, we have

$$\mathbb{P}\left(\left|\tilde{V}_1'(n,m) - F(n,k)\right| \geq \sqrt{nk}\epsilon\right) \leq C_4 k\sqrt{np_{max}^m} \exp\left\{-\frac{\sqrt{np_{max}^m}\epsilon}{C_4\sqrt{k}}\right\}, \quad (1.3.16)$$

for some positive constant C_4. From (1.3.12), for any $x > 2$ and $0 < \epsilon < (x-2)$,

$$\mathbb{P}(F(n,k) \geq \sqrt{nk}(x \pm \epsilon)) = \exp\{-k(I_1(x \pm \epsilon) + o(1))\}. \quad (1.3.17)$$

Hence,

$$\begin{aligned}
&\frac{\mathbb{P}\left(\left|\tilde{V}_1'(n,m) - F(n,k)\right| \geq \sqrt{nk}\epsilon\right)}{\mathbb{P}(F(n,k) \geq \sqrt{nk}(x \pm \epsilon))} \\
&\leq C_4 k\sqrt{np_{max}^m} \exp\left\{\sqrt{\frac{np_{max}^m}{k}}\left[-\frac{\epsilon}{C_4} + \sqrt{\frac{k^3}{np_{max}^m}}\left(I_1(x \pm \epsilon) + o(1)\right)\right]\right\} \\
&\to 0, \quad \text{as } n \to \infty, \quad k(m(n))^3/p_{max}^m = o(n),
\end{aligned}$$

and as in the proof of Theorem 1.1.1, this leads to (1.3.4) for any $x > 2$. Applying the same arguments at the end of the proof of Theorem 1.1.1 we can prove that (1.3.4) is valid for any $x \geq 2$.

The proof of (1.3.5) is similar to the uniform case. First, from (1.3.10) and (1.3.11), for any fixed $x < 2$,

$$\lim_{k\to\infty} \frac{1}{k} \log \mathbb{P}(F(1,k) \leq \sqrt{k}x) = -\infty. \tag{1.3.18}$$

Moreover, for any $0 < \epsilon < 2 - x$,

$$\mathbb{P}\left(\tilde{V}_1^{'}(n,m) \leq \sqrt{nk}x\right) \leq \\ \mathbb{P}\left(F(n,k) \leq \sqrt{nk}(x+\epsilon)\right) + \mathbb{P}\left(\left|\tilde{V}_1^{'}(n,m) - F(n,k)\right| \geq \sqrt{nk}\epsilon\right), \tag{1.3.19}$$

while $\mathbb{P}\left(\left|\tilde{V}_1^{'}(n,m) - F(n,k)\right| \geq \sqrt{nk}\epsilon\right)$ can be further controlled by $e^{-k(m)T}$, with $T > 0$, arbitrarily large. Hence (1.3.5) holds true under the condition $k(m(n))^3/p_{max}^m = o(n)$. ∎

Proof of Theorem 1.1.3

Set $X = (V_1(n,m) - np_{max}^m)/\sqrt{nkp_{max}^m}$, $Y = (V_1(n,m) - V_1^{'}(n,m))/\sqrt{nkp_{max}^m}$ and $Z = (V_1^{'}(n,m) - np_{max}^m)/\sqrt{nkp_{max}^m}$. For any $x > 2$ and $0 < \epsilon < x - 2$,

$$\mathbb{P}(X \geq x) \leq \mathbb{P}(Z \geq x - \epsilon) + \mathbb{P}(|Y| \geq \epsilon), \tag{1.3.20}$$

and

$$\mathbb{P}(X \geq x) \geq \mathbb{P}(Z \geq x + \epsilon) - \mathbb{P}(|Y| \geq \epsilon). \tag{1.3.21}$$

Moreover, from (1.3.1)

$$\mathbb{P}(|Y| \geq \epsilon) \leq \frac{Cp_{2nd}^m\sqrt{n}}{\epsilon\sqrt{kp_{max}^m}}, \tag{1.3.22}$$

and from Lemma 2.3.2,

$$\mathbb{P}(Z \geq x \pm \epsilon) = \exp\{-k[I_1(x \pm \epsilon) + o(1)]\}.$$

Under the condition (1.1.15), we have

$$\frac{\mathbb{P}(|Y| \geq \epsilon)}{\mathbb{P}(Z \geq x \pm \epsilon)} \leq \frac{C p_{2nd}^m \sqrt{n}}{\epsilon \sqrt{k p_{max}^m}} \exp\{k[I_1(x \pm \epsilon) + o(1)]\} \to 0, \quad \text{as } n \to \infty. \tag{1.3.23}$$

Letting ϵ go to 0, and repeating the same arguments as in the proof of Theorem 1.1.1, proves (1.1.16), for any $x \geq 2$, under the conditions given in Theorem 1.1.3.
For (1.1.17), for any $x < 2$ and $0 < \epsilon < 2 - x$,

$$\mathbb{P}(X \leq x) \leq \mathbb{P}(Z \leq x + \epsilon) + \mathbb{P}(|Y| \geq \epsilon).$$

From (1.3.3), $\mathbb{P}(Z \leq x + \epsilon)$ is exponentially negligible with speed $k(m)$, and from arguments as in (1.3.23), $\mathbb{P}(|Y| \geq \epsilon)$ is bounded by $e^{-k(m)T}$, $T > 0$, as $n \to \infty$. Hence, letting $T \to \infty$, $\mathbb{P}(|Y| \geq \epsilon)$ is also exponentially negligible with speed $k(m)$, which proves (1.1.17). ■

Proof of Theorem 1.1.4 and Remark 1.1.3

First, (1.1.22) is a direct consequence of (1.1.16). Next, we prove (1.1.21). As in the proof of Lemma 2.3.2, when $V_1'(n, m)$, is on the left of its simultaneous asymptotic mean, it can be approximated by $F(n, k)$ (see (1.3.8)). Hence the rate function K_η should be the corresponding rate function of the Brownian functional $F(1, k)$ (see (1.3.9)) when it is on the left of its asymptotic mean, with convergence rate $k(m)^2$. From the right hand side of (1.3.10) we know that this new rate function K_η will depend on η, which is the limit of $k p_{max}^m$. Moreover, for $F(1, k)$, and from [22],

$$\sqrt{1 - p_{max}^m} F(1, k) \stackrel{\mathcal{L}}{=} \tilde{\lambda}_1^0,$$

where $\tilde{\lambda}_1^0$ is the largest eigenvalue of the diagonal block corresponding to p_{max}^m in $\mathbf{X^0}$, and where $\mathbf{X^0}$ is an element of $\mathcal{G}^0(p_1^m, p_2^m, \cdots, p_m^m)$. So the rate function K_η should also be the corresponding rate function of $\tilde{\lambda}_1^0$ when it is on the left of its asymptotic mean with convergence rate $k(m)^2$.

Again, from [22],

$$\lambda_1^k \stackrel{\mathcal{L}}{=} \tilde{\lambda}_1^0 + \sqrt{p_{max}^m} g. \tag{1.3.24}$$

where λ_1^k is the largest eigenvalue of the $k \times k$ GUE and g is a standard normal random variable which is independent of $\tilde{\lambda}_1^0$.

Let

$$J(x) = \begin{cases} \inf_{\mu \in \mathcal{M}((-\infty, x])} I(\mu), & \text{if } x \in (-\infty, 2], \\ 0, & \text{if } x \in [2, +\infty), \end{cases} \tag{1.3.25}$$

$$G_\eta(x) = \begin{cases} \frac{x^2}{2\eta}, & \text{if } x \in (-\infty, 0], \\ 0, & \text{if } x \in [0, +\infty), \end{cases} \tag{1.3.26}$$

where J is the rate function for λ_1^k with speed k^2, $I(\mu)$ is given in (1.5.5), and G_η is the corresponding rate function for the Gaussian term. Now, see [15], when $x \leq 2$,

$$J(x) = \frac{1}{216}\Bigg(- x\left(-72x + x^3 + 30\sqrt{12 + x^2} + x^2\sqrt{12 + x^2}\right) \\ - 216 \log\left(\frac{1}{6}\left(x + \sqrt{12 + x^2}\right)\right)\Bigg). \tag{1.3.27}$$

Hence,

$$J'(x) = \frac{1}{54}\left(-x^3 + 36x - (12 + x^2)^{3/2}\right), \tag{1.3.28}$$

$$J''(x) = \frac{1}{18}\left(12 - x^2 - x\sqrt{12 + x^2}\right). \tag{1.3.29}$$

Notice that $0 < J''(x) < 1$ for $x \in (-\infty, 2)$. Moreover, by Taylor expansions for J and J', and for $x < -5$,

$$J(x) = \frac{x^2}{2} + \log(-x) + \frac{3}{4} + e_1(x), \tag{1.3.30}$$

$$J'(x) = x + \frac{1}{x} + e_2(x), \tag{1.3.31}$$

with $|e_1(x)| \leq 2/x^2$ and $|e_2(x)| \leq 4/|x|^3$.

From (1.3.24), it is well known (see [16], [34]) that,

$$J(x) = K_\eta \square G_\eta(x) := \inf_{y \in \mathbb{R}} \{K_\eta(y) + G_\eta(x - y)\}, \tag{1.3.32}$$

and taking Legendre transforms to get

$$K_\eta(x) = \left(J^*(y) - G_\eta^*(y)\right)^*(x),$$

where

$$G^*(y) = \begin{cases} \frac{\eta y^2}{2}, & \text{if } y \leq 0, \\ +\infty, & \text{if } y > 0, \end{cases}$$

so

$$K_\eta(x) = \sup_{y \leq 0} \left(xy - J^*(y) + \frac{\eta y^2}{2} \right). \tag{1.3.33}$$

Hence for $\eta = 0$, $K_0 = J$, for $\eta = 1$, $K_1 = K$, while for $0 < \eta < 1$, K_η interpolates between J and K. From the very definition of the Legendre transform,

$$J^*(y) = \sup_{x \in \mathbb{R}} \left(xy - J(x) \right),$$

for each $y \leq 0$, there exists a unique solution to $J'(x) = y$ on $x \in (-\infty, 2]$, and we denote this solution by $S(y)$. S is an increasing function on $(-\infty, 0]$ with $S(0) = 2$, $\lim_{y \to -\infty} S(y) = -\infty$ and

$$S'(y) = \frac{1}{J''(S(y))},$$

for $y < 0$. Thus for $y \leq 2$,

$$J^*(y) = yS(y) - J(S(y)).$$

And as a consequence,

$$K_\eta(x) = \sup_{y \leq 0} \left(xy - yS(y) + J(S(y)) + \frac{\eta y^2}{2} \right).$$

For $y \leq 0$, let

$$H_{x,\eta}(y) := xy - yS(y) + J(S(y)) + \frac{\eta y^2}{2},$$

then

$$H'_{x,\eta}(y) = x - S(y) + \eta y, \;\; H''_{x,\eta}(y) = -\frac{1}{J''(S(y))} + \eta,$$

so $H''_{x,\eta}(y) < 0$ for $y \in (-\infty, 0)$, $x \in \mathbb{R}$ and $0 \le \eta \le 1$. When $x \ge 2$, for any $0 \le \eta \le 1$, $H'_{x,\eta}(y) > 0$ for $y < 0$ with $H'_{x,\eta}(0) \ge 0$, thus $K_\eta(x) = \sup_{y \le 0} H_{x,\eta}(y) = H_{x,\eta}(0) = 0$.

Now we consider the case when $x < 2$. First, from (1.3.31), it can be shown that for $y < -6$,

$$y < S(y) < y + 1,$$

and thus since $x - J'(x)$ is increasing on $(-\infty, 2]$,

$$S(y) - y = S(y) - J'(S(y)) < y + 1 - J'(y + 1) < -\frac{2}{y+1},$$

which further yields

$$y < S(y) < y - \frac{2}{y+1}.$$

Moreover, when $y < -6$,

$$\begin{aligned} \left| H_{x,\eta}(y) - \left(xy - y^2 + J(y) + \frac{\eta y^2}{2} \right) \right| &\le |y|\,|S(y) - y| + |J(S(y)) - J(y)| \\ &\le 2\left| \frac{y}{y+1} \right| + |J'(y)|\,|S(y) - y| \\ &\le 3 + 3 = 6. \end{aligned} \tag{1.3.34}$$

Combining (1.3.34) with (1.3.30), we get that for $y < -6$,

$$\left| H_{x,\eta}(y) - \left(xy + \log(-y) - \frac{1-\eta}{2} y^2 \right) \right| \le 7. \tag{1.3.35}$$

When $\eta = 1$, for any $x \le 0$, $H'_{x,1}(y) < 0$ for $y \le 0$, thus

$$K_1(x) = \lim_{y \to -\infty} H_{x,1}(y) = +\infty.$$

For $0 < x < 2$, since $S(y) - y$ is increasing on $(-\infty, 0]$ with a range of $(0, 2]$, there exists a unique solution to $H'_{x,1}(y) = x - S(y) + y = 0$, and we denote it by $T_1(x)$. Note that $y = T_1(x)$ is the maximizer for $H_{x,1}(y)$ and as $x \to 0$, $T_1(x) \to -\infty$, thus there exists some $\delta > 0$, such that when $x < \delta$,

$$K_1(x) = \sup_{y \le -6} H_{x,1}(y).$$

Since for $x < 1/6$,

$$\sup_{y \leq -6} (xy + \log(-y)) = -1 - \log x,$$

combining this with (1.3.35) gives for x close enough to 0,

$$|K_1(x) - (-\log x)| \leq 8.$$

When $0 < \eta < 1$, for any $x < 2$, there exists a unique solution to $H'_{x,\eta}(y) = x - S(y) + \eta y = 0$, which is denoted by $T_\eta(x)$. Note that $y = T_\eta(x)$ is the maximizer of $H_{x,\eta}(y)$ and as $x \to -\infty$, $T_\eta(x) \to -\infty$. By repeating arguments as in the case $\eta = 1$ we get as $x \to -\infty$,

$$K_\eta(x) \sim \frac{x^2}{2(1-\eta)} + \log\left(-\frac{x}{1-\eta}\right),$$

which is consistent with $J(x)$ when $\eta = 0$.

The rest of the proof follows exactly the proof of Lemma 2.3.2 and of Theorem 1.1.3. ∎

1.4 Proof of Theorem 1.1.5 and Theorem 1.1.6

Left and right concentration inequalities for the largest eigenvalue λ_1^m of an element of the $m \times m$ GUE are respectively given in [3] and [27]. More precisely:

Proposition 1.4.1 *Let $m \geq 1$ and let $\epsilon > 0$, then for some absolute positive constant C_0,*

$$\mathbb{P}(\lambda_1^m \geq 2\sqrt{m}(1+\epsilon)) \leq C_0 e^{-m\epsilon^{3/2}/C_0}. \tag{1.4.1}$$

Likewise, for some absolute positive constant $\bar{C}_0$, and all $m \geq 1$ and $0 < \epsilon \leq 1$,

$$\mathbb{P}(\lambda_1^m \leq 2\sqrt{m}(1-\epsilon)) \leq \bar{C}_0 e^{-m^2\epsilon^3/\bar{C}_0}. \tag{1.4.2}$$

Next, to prove (1.1.24), assume first that $m\epsilon^{3/2} \geq 1$. Then for any $0 < \epsilon < 1$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{V_1(n,m) - n/m}{\sqrt{n/m}} \geq 2\sqrt{m}(1+\epsilon)\right) \\
&\leq \mathbb{P}\left(\sqrt{\frac{m-1}{m}}\frac{\tilde{L}_1(n,m)}{2\sqrt{mn}} \geq 1 + \frac{\epsilon}{2}\right) \\
&\qquad + \mathbb{P}\left(\sqrt{\frac{m-1}{m}}\frac{|\tilde{V}_1(n,m) - \tilde{L}_1(n,m)|}{2\sqrt{mn}} \geq \frac{\epsilon}{2}\right). \qquad (1.4.3)
\end{aligned}$$

As previously,

$$\sqrt{\frac{m-1}{m}}\frac{\tilde{L}_1(n,m)}{\sqrt{n}} \stackrel{\mathcal{L}}{=} \lambda_1^{m,0},$$

and

$$\lambda_1^m \stackrel{\mathcal{L}}{=} \lambda_1^{m,0} + Z_m,$$

where Z_m is a centered Gaussian random variable with variance $1/m$, which is independent of $\lambda_1^{m,0}$. So,

$$\begin{aligned}
\mathbb{P}\left(\sqrt{\frac{m-1}{m}}\frac{\tilde{L}_1(n,m)}{2\sqrt{mn}} \geq 1 + \frac{\epsilon}{2}\right) &\leq \mathbb{P}\left(\lambda_1^m \geq 2\sqrt{m}\left(1 + \frac{\epsilon}{4}\right)\right) + \mathbb{P}\left(Z_m \geq \frac{\sqrt{m}\epsilon}{2}\right) \\
&\leq C_1 e^{-m\epsilon^{3/2}/C_1} + C_1 e^{-m^2\epsilon^2/C_1},
\end{aligned}$$

for some positive constant C_1. Now from (1.2.7), (1.2.8) and (1.2.9), the second term on the right hand side of (1.4.3) is bounded by:

$$\mathbb{P}\left(\frac{|\tilde{V}_1(n,m) - \tilde{L}_1(n,m)|}{2\sqrt{mn}} \geq \frac{\epsilon}{2}\right) \leq C_2\sqrt{mn}e^{-\sqrt{n}\epsilon/C_2 m} + C_2 mne^{-n\epsilon^2/C_2 m}.$$

In order to reach (1.1.24), we need to show that there exists a positive constant $C(A,\alpha)$, depending only on A and α, such that

$$C(A,\alpha)e^{-m\epsilon^{3/2}/C(A,\alpha)} \geq C_1 e^{-m^2\epsilon^2/C_1}, \qquad (1.4.4)$$

$$C(A,\alpha)e^{-m\epsilon^{3/2}/C(A,\alpha)} \geq C_2\sqrt{mn}e^{-\sqrt{n}\epsilon/C_2 m}, \qquad (1.4.5)$$

$$C(A,\alpha)e^{-m\epsilon^{3/2}/C(A,\alpha)} \geq C_2 mne^{-n\epsilon^2/C_2 m}. \qquad (1.4.6)$$

First, since $m\epsilon^{3/2} \geq 1$, (1.4.4) can be satisfied by choosing $C(A,\alpha) \geq C_1$. Now taking logarithms in (1.4.5), $C(A,\alpha)$ has to be such that:

$$\log \frac{C_2}{C(A,\alpha)} + \frac{1}{2}\log(mn) \leq m\epsilon^{3/2}\left(-\frac{1}{C(A,\alpha)} + \frac{\sqrt{n}}{C_2 m^2 \epsilon^{1/2}}\right). \tag{1.4.7}$$

Moreover, under the condition $m \leq An^{\alpha}$, we have:

$$\frac{\sqrt{n}}{C_2 m^2 \epsilon^{1/2}} \geq \frac{\sqrt{n}}{C_2 m^2} \geq \frac{n^{\frac{1}{2}-2\alpha}}{A^2 C_2}.$$

Therefore, if $\alpha < 1/4$, we just need to choose $C(A,\alpha)$ satisfying

$$\log \frac{\sqrt{A}C_2}{C(A,\alpha)} + \frac{1}{C(A,\alpha)} \leq \frac{n^{\frac{1}{2}-2\alpha}}{A^2 C_2} - \frac{1+\alpha}{2}\log n.$$

Since for all integers $n \geq 1$,

$$\frac{n^{\frac{1}{2}-2\alpha}}{A^2 C_2} - \frac{1+\alpha}{2}\log n \geq \frac{1+\alpha}{1-4\alpha}\left(1 - \log\frac{A^2 C_2(1+\alpha)}{1-4\alpha}\right),$$

we just need to guarantee that

$$\log \frac{\sqrt{A}C_2}{C(A,\alpha)} + \frac{1}{C(A,\alpha)} \leq \frac{1+\alpha}{1-4\alpha}\left(1 - \log\frac{A^2 C_2(1+\alpha)}{1-4\alpha}\right). \tag{1.4.8}$$

But from our choice of α, $(1+\alpha)/(1-4\alpha) > 1$, so by choosing

$$C(A,\alpha) \geq C\max\{A^{5/2}, 1\}\frac{1+\alpha}{1-4\alpha}\exp\left\{\frac{1+\alpha}{1-4\alpha}\right\}, \tag{1.4.9}$$

for some large enough absolute constant C, (1.4.8) and (1.4.5) are satisfied.

Finally, by taking logarithms, (1.4.6) becomes,

$$\log \frac{C_2}{C(A,\alpha)} + \log(mn) \leq m\epsilon^{3/2}\left(-\frac{1}{C(A,\alpha)} + \frac{n\epsilon^{1/2}}{C_2 m^2}\right). \tag{1.4.10}$$

From the condition $m \leq An^{\alpha}$, we just need,

$$\log \frac{AC_2}{C(A,\alpha)} + \frac{1}{C(A,\alpha)} \leq \frac{1}{A^{7/3}C_2} n^{1-\frac{7\alpha}{3}} - (1+\alpha)\log n. \tag{1.4.11}$$

Now repeating the previous arguments, taking the minimum on the right hand side of (1.4.11), we have

$$\log \frac{AC_2}{C(A,\alpha)} + \frac{1}{C(A,\alpha)} \leq \frac{1+\alpha}{1-7\alpha/3}\left(1 - \log\frac{A^{7/3}C_2(1+\alpha)}{1-7\alpha/3}\right). \tag{1.4.12}$$

Again, for $0 < \alpha < 1/4$, $1 < (1+\alpha)/(1-7\alpha/3) < 3$, so as long as we choose

$$C(A,\alpha) \geq C \max\{A^{10/3}, 1\} \frac{1+\alpha}{1-7\alpha/3} \exp\left\{\frac{1+\alpha}{1-7\alpha/3}\right\}, \tag{1.4.13}$$

for some large enough absolute constant C, $C(A,\alpha)$ will satisfy (1.4.12) and hence also satisfy (1.4.6).

Combining (1.4.9) and (1.4.13), if $m\epsilon^{3/2} \geq 1$, and $m \leq An^\alpha$, with $\alpha < 1/4$, we can find a positive constant

$$C(A,\alpha) = C \max\{A^{10/3}, 1\} \frac{1+\alpha}{1-4\alpha} \exp\left\{\frac{1+\alpha}{1-4\alpha}\right\}, \tag{1.4.14}$$

so that (1.1.24) holds for all $0 < \epsilon < 1$. When $m\epsilon^{3/2} < 1$,

$$C(A,\alpha) e^{-m\epsilon^{3/2}/C(A,\alpha)} \geq Ce^{-1/C} \geq 1,$$

as C is large enough, and (1.1.24) follows naturally. So combining these two cases, we can find a positive $C(A,\alpha)$ as in (1.4.14), with C large enough, such that (1.1.24) holds.

Likewise, for the proof of (1.1.25), first assume that $m^2\epsilon^3 \geq 1$, and

$$\begin{aligned}
&\mathbb{P}\left(\frac{V_1(n,m) - n/m}{\sqrt{n/m}} \geq 2\sqrt{m}(1-\epsilon)\right) \qquad (1.4.15)\\
&\leq \mathbb{P}\left(\sqrt{\frac{m-1}{m}}\frac{\tilde{L}_1(n,m)}{2\sqrt{mn}} \leq 1 - \frac{\epsilon}{2}\right)\\
&\qquad + \mathbb{P}\left(\sqrt{\frac{m-1}{m}}\frac{|\tilde{V}_1(n,m) - \tilde{L}_1(n,m)|}{2\sqrt{mn}} \geq \frac{\epsilon}{2}\right)\\
&\leq C_1 e^{-m^2\epsilon^3/C_1} + C_1 e^{-m^2\epsilon^2/C_1} + C_2\sqrt{mn}e^{-\sqrt{n}\epsilon/C_2 m} + C_2 mne^{-n\epsilon^2/C_2 m}.
\end{aligned}$$

Repeating previous arguments, we get that, as long as $m \leq An^\alpha$, with $\alpha < 1/6$, we can find some positive constant

$$\bar{C}(A,\alpha) = \bar{C} \max\{A^4, 1\} \frac{1+\alpha}{1-6\alpha} \exp\left\{\frac{1+\alpha}{1-6\alpha}\right\},$$

so that (1.1.25) is satisfied. Again, by taking $\bar{C}$ large enough, the case $m^2\epsilon^3 < 1$ follows, and (1.1.25) is proved. ■

The proof for the non-uniform case is similar to the uniform one. For (1.1.27), we first assume that $k\epsilon^{3/2} \geq 1$, then

$$\begin{aligned}
&\mathbb{P}\left(\frac{V_1(n,m) - np_{max}^m}{\sqrt{nkp_{max}^m}} \geq 2(1+\epsilon)\right) \\
&\leq \mathbb{P}\left(\frac{V_1(n,m) - V_1^{'}(n,m)}{2\sqrt{nkp_{max}^m}} \geq \frac{\epsilon}{3}\right) + \mathbb{P}\left(\sqrt{1-p_{max}^m}\frac{\tilde{V}_1^{'}(n,m) - F(n,k)}{2\sqrt{nk}} \geq \frac{\epsilon}{3}\right) \\
&+ \mathbb{P}\left(\sqrt{1-p_{max}^m}\frac{F(n,k)}{2\sqrt{nk}} \geq 1 + \frac{\epsilon}{3}\right) \\
&= A_1 + A_2 + A_3.
\end{aligned}$$

From (1.3.22), (1.3.13) and (1.3.10), we have

$$\begin{aligned}
A_1 &\leq \frac{C_1 p_{2nd}^m \sqrt{n}}{\epsilon\sqrt{kp_{max}^m}}, \\
A_2 &\leq C_2 nk \exp\left\{-\frac{n\epsilon^2}{C_2 k}\right\} + C_2 k\sqrt{np_{max}^m}\exp\left\{-\frac{\sqrt{np_{max}^m}\epsilon}{C_2\sqrt{k}}\right\}, \\
A_3 &\leq \mathbb{P}\left(Z_k \geq \frac{\epsilon}{3}\right) + \mathbb{P}\left(\lambda_1^k \geq 2\left(1+\frac{\epsilon}{6}\right)\right) \\
&\leq C_3 \exp\left\{-\frac{k^2\epsilon^2}{C_3}\right\} + C_3 \exp\left\{-\frac{k\epsilon^{3/2}}{C_3}\right\}.
\end{aligned}$$

In order to reach (1.1.27), we need to show that there exists a positive $C(A,B,\alpha)$, depending only on A, B and α, such that

$$C(A,B,\alpha)\exp\left\{-\frac{k\epsilon^{3/2}}{C(A,B,\alpha)}\right\} \geq \frac{C_1 p_{2nd}^m\sqrt{n}}{\epsilon\sqrt{kp_{max}^m}}, \tag{1.4.16}$$

$$C(A,B,\alpha)\exp\left\{-\frac{k\epsilon^{3/2}}{C(A,B,\alpha)}\right\} \geq C_2 nk\exp\left\{-\frac{n\epsilon^2}{C_2 k}\right\}, \tag{1.4.17}$$

$$C(A,B,\alpha)\exp\left\{-\frac{k\epsilon^{3/2}}{C(A,B,\alpha)}\right\} \geq C_2 k\sqrt{np_{max}^m}\exp\left\{-\frac{\sqrt{np_{max}^m}\epsilon}{C_2\sqrt{k}}\right\}, \tag{1.4.18}$$

$$C(A,B,\alpha)\exp\left\{-\frac{k\epsilon^{3/2}}{C(A,B,\alpha)}\right\} \geq C_3\exp\left\{-\frac{k^2\epsilon^2}{C_3}\right\}. \tag{1.4.19}$$

First, by taking logarithms in (1.4.18), we get

$$\log\frac{C_2}{C(A,B,\alpha)} + \log k + \frac{1}{2}\log(np_{max}^m) \leq k\epsilon^{3/2}\left(-\frac{1}{C(A,B,\alpha)} + \frac{\sqrt{np_{max}^m}}{C_2\sqrt{\epsilon k^3}}\right).$$

Next,

$$\frac{\sqrt{np_{max}^m}}{C_2\sqrt{\epsilon k^3}} \geq \frac{\sqrt{(np_{max}^m)^{1-3/\alpha}}}{A^{3/2\alpha}C_2},$$

so if $\alpha > 3$, then we can choose a constant $C(A, B, \alpha)$, satisfying (1.4.18). Actually here $C(A, B, \alpha)$ just needs to satisfy

$$\log \frac{A^{1/\alpha} C_2}{C(A, B, \alpha)} + \frac{1}{C(A, B, \alpha)} \leq \frac{\alpha + 2}{\alpha - 3} \left(1 - \log \frac{A^{3/2\alpha} C_2 (\alpha + 2)}{\alpha - 3} \right),$$

which forces

$$C(A, B, \alpha) \geq C \max\{A^{2/\alpha}, 1\} \frac{\alpha + 2}{\alpha - 3} \exp\left\{ \frac{\alpha + 2}{\alpha - 3} \right\}, \tag{1.4.20}$$

for a large enough absolute constant C.

Second, by taking logarithms in (1.4.16), we have:

$$\log \frac{C_1}{C(A, B, \alpha)} + \log \left(\frac{p_{2nd}^m \sqrt{n}}{\sqrt{k p_{max}^m}} \right) \leq - \frac{k \epsilon^{3/2}}{C(A, B, \alpha)} + \log \epsilon.$$

From (1.1.26) and the assumption $k\epsilon^{3/2} \geq 1$, in order for (1.4.16) to hold true, $C(A, B, \alpha)$ needs to satisfy

$$\log \frac{C_1 \sqrt{B}}{C(A, B, \alpha)} - \frac{k}{2} \leq - \frac{k}{C(A, B, \alpha)} - \frac{2}{3} \log k,$$

which further forces

$$C(A, B, \alpha) \geq C \max\{\sqrt{B}, 1\}, \tag{1.4.21}$$

with the absolute constant C large enough.

For (1.4.17), as we did in (1.4.6), and under the condition $k^\alpha / p_{max}^m \leq An$ with $\alpha > 3$, we need to choose

$$C(A, B, \alpha) \geq C \max\{A^{10/3\alpha}, 1\} \frac{3\alpha + 3}{3\alpha - 7} \exp\left\{ \frac{3\alpha + 3}{3\alpha - 7} \right\}, \tag{1.4.22}$$

with the absolute constant C large enough. Finally, (1.4.19) is easy to satisfy since $k\epsilon^{3/2} \geq 1$. Moreover, when $k\epsilon^{3/2} < 1$, then (1.1.27) holds naturally given C large enough.

Combining (1.4.20), (1.4.21) and (1.4.22), choosing

$$C(A, B, \alpha) = C \max\{A^{10/3\alpha}, 1\} \max\{\sqrt{B}, 1\} \frac{\alpha + 2}{\alpha - 3} \exp\left\{ \frac{\alpha + 2}{\alpha - 3} \right\},$$

with C some large enough absolute constant, (1.1.27) holds under the given conditions. Likewise, we can prove (1.1.29). ∎

1.5 Large deviations for the spectrum of the traceless GUE

For any integer $m \geq 2$, let the random matrix $\mathbf{X}$ be an element of the $m \times m$ GUE. Let $(\lambda_1, \lambda_2, \cdots, \lambda_m)$ be the spectrum of $\mathbf{X}$, and let

$$(\xi_1, \xi_2, \cdots, \xi_m) = \frac{1}{\sqrt{m}}(\lambda_1, \lambda_2, \cdots, \lambda_m).$$

The joint probability density of $(\xi_1, \xi_2, \cdots, \xi_m)$ is given by

$$\phi_m(\xi_1, \xi_2, \cdots, \xi_m) = \frac{1}{Z_m} \exp\left\{-\frac{m}{2}\sum_{i=1}^{m} \lambda_i^2\right\} \prod_{1 \leq i < j \leq m} (\lambda_i - \lambda_j)^2, \tag{1.5.1}$$

where

$$Z_m = (2\pi)^{\frac{m}{2}} m^{-\frac{m^2}{2}} \prod_{j=1}^{m} j!, \tag{1.5.2}$$

see Theorem 2.5.2 in [5] and also Theorem 3.3.1 in [32].

Let $(\lambda_1^0, \lambda_2^0, \cdots, \lambda_m^0)$ be the spectrum of $\mathbf{X} - tr(\mathbf{X})/m$, an element of the $m \times m$ traceless GUE, and again, let

$$(\xi_1^0, \xi_2^0, \cdots, \xi_m^0) = \frac{1}{\sqrt{m}}(\lambda_1^0, \lambda_2^0, \cdots, \lambda_m^0).$$

The joint distribution function of $(\xi_1^0, \xi_2^0, \cdots, \xi_m^0)$ is given by

$$\begin{aligned} &\mathbb{P}\left(\xi_1^0 \leq s_1, \xi_2^0 \leq s_2, \cdots, \xi_m^0 \leq s_m\right) \\ &= \sqrt{2\pi} \int_{\mathcal{L}(s_1, \cdots, s_m)} \phi_m(x_1, x_2, \cdots, x_m) dx_1 \cdots dx_{m-1}, \end{aligned} \tag{1.5.3}$$

where

$$\mathcal{L}(s_1, \cdots, s_m) := \Bigg\{ x = (x_1, \cdots, x_m) \in \mathbb{R}^m : \sum_{i=1}^{m} x_i = 0, \textit{and } x_i < s_j, \textit{ for each } i = 1, \cdots, m \Bigg\}.$$

Let $(\xi_1^m, \xi_2^m, \cdots, \xi_m^m)$ be the nonincreasing rearrangement of $(\xi_1, \xi_2, \cdots, \xi_m)$, and let $(\xi_1^{m,0}, \xi_2^{m,0}, \cdots, \xi_m^{m,0})$ be the nonincreasing rearrangement of $(\xi_1^0, \xi_2^0, \cdots, \xi_m^0)$, then, *e.g.*, see [22],

$$(\xi_1^m, \xi_2^m, \cdots, \xi_m^m) \stackrel{\mathcal{L}}{=} (\xi_1^{m,0}, \xi_2^{m,0}, \cdots, \xi_m^{m,0}) + g_m \mathbf{e}_m, \tag{1.5.4}$$

where g_m is a centered Gaussian random variable with variance $1/m^2$, independent of the vector $(\xi_1^{m,0}, \xi_2^{m,0}, \cdots, \xi_m^{m,0})$, and where $\mathbf{e}_m = (1, 1, \cdots, 1)$.

As shown in [10], the law of the spectral measure $\hat{\mu}^m = \frac{1}{m}\sum_{i=1}^m \delta_{\xi_i}$ satisfies a large deviation principle on the set $\mathcal{P}(\mathbb{R})$ of probability measures on $\mathbb{R}$, and with good rate function I, in the scale m^2. Moreover, I is given by

$$I(\mu) = \frac{1}{2}\int x^2 d\mu(x) - \iint \log|x-y| d\mu(x) d\mu(y) - \frac{3}{4}, \tag{1.5.5}$$

and its unique minimizer is the semicircular probability measure

$$\sigma = \frac{1}{2\pi}\mathbf{1}_{|x|\le 2}\sqrt{4-x^2}dx.$$

Based on this LDP for $\hat{\mu}^m$, the LDP for the largest (or rth largest) eigenvalue of the GOE with an explicit rate function is obtained in [9] and [4] (see also [24] for generalizations). Following the approach and the techniques developed there, we can prove the multidimensional LDP for the first r eigenvalues of the GUE:

Theorem 1.5.1 *Let $r \in \mathbb{N}$, on $\mathcal{L}^r := \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \ge x_2 \ge \cdots \ge x_r\}$, $(\xi_1^m, \xi_2^m, \cdots, \xi_r^m)$ satisfies a LDP with speed m and a good rate function*

$$I_r(x_1, x_2, \cdots, x_r) = \begin{cases} 2\sum_{i=1}^r \int_2^{x_i} \sqrt{(z/2)^2 - 1}dz, & \text{if } x_1 \ge x_2 \cdots \ge x_r \ge 2, \\ +\infty, & \text{otherwise.} \end{cases}$$

Proof. First, for the convenience of the future proof, we may denote the joint density of the spectrum of the $m \times m$ GUE as

$$Q_m(d\xi_1, d\xi_2, \cdots, d\xi_m) = \frac{1}{Z_m}\exp\left\{-\frac{m}{2}\sum_{i=1}^m \xi_i^2\right\}\prod_{1\le i<j\le m}(\xi_i - \xi_j)^2 \prod_{i=1}^m d\xi_i.$$

To prove this theorem, we need the lemma:

Lemma 1.5.1 *For every positive number L large enough and all m,*

$$Q_m\left(\max_{i=1}^m |\xi_i| \ge L\right) \le e^{-mL^2/5}.$$

Proof of Lemma 1.5.1:

Observe that since L is large enough, for any $|x| \geq L$ and $\xi_i \in \mathbb{R}$,

$$|x - \xi_i|^2 e^{-\xi_i^2/2} \leq 2(|x|^2 + |\xi_i|^2)e^{-\xi_i^2/2} \leq 4|x|^2 \leq e^{x^2/4}$$

So we have:

$$\begin{aligned} Q_m(|\xi_1| \geq L) &\leq e^{-\frac{1}{4}mL^2} \frac{Z_{m-1}}{Z_m} \int_{|x|\geq L} e^{-x^2/4} dx \\ &\quad \times \int \prod_{i=2}^{m} (|x - \xi_i|^2 e^{-\xi_i^2/2} e^{-x^2/4}) Q_{m-1}(d\xi_2, \cdots, \xi_m) \\ &\leq e^{-\frac{1}{4}mL^2} \frac{Z_{m-1}}{Z_m} \int e^{-x^2/4} dx \end{aligned}$$

Further, the explicit formula for Z_m shows that $Z_{m-1}/Z_m \leq e^{C'm}$ for some finite C' and all m. Taking $C = \max(C', \int e^{-x^2/4} dx)$, we have:

$$Q_m \left(\max_{i=1}^{m} |\xi_i| \geq L \right) \leq m Q_m(|\xi_1| \geq L) \leq e^{-\frac{1}{4}mL^2 + 2Cm}$$

and the lemma follows since $C < \infty$ is independent of L. ∎

Now we return to the proof of Theorem 1.5.1. To show Theorem 1.5.1 it is sufficient to prove

$$\limsup_{m\to\infty} \frac{1}{m} \log Q_m(\xi_r^m \leq x) = -\infty \quad \text{for all } x < 2, \tag{1.5.6}$$

and, since $I_r(x_1, x_2, \cdots, x_r)$ is continuous on $\mathcal{L}^r \cap [2, \infty)^r$,

$$\lim_{m\to\infty} \frac{1}{m} \log Q_m \left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r\right) = -2 \sum_{i=1}^{r} \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz$$
$$\text{for all } x_1 \geq x_2 \cdots \geq x_r \geq 2 \tag{1.5.7}$$

First, suppose that $\xi_r^m \leq x$ for some $x < 2$, then we have $\hat{\mu}^m((x, 2]) \leq (r-1)/m$. Since $\sigma((x, 2]) > 0$, we may find a bounded and continuous function $h(y)$ such that $h(y) = 0$ for all $y \leq x$ and $\int h d\sigma > 0$. Choose some number r such that $0 < r < \int h d\sigma$ and define a closed set of $\mathcal{P}(\mathbb{R})$ as $F := \{\nu \in \mathcal{P}(\mathbb{R}) : \int h d\nu \leq r\}$, then $\sigma \notin F$ but for m large enough, $\hat{\mu}^m \in F$. By the LDP for $\hat{\mu}^m$, we know $Q_m(\hat{\mu}^m \in F) \leq e^{-cm^2}$ for

some $c > 0$, so (1.5.6) follows consequently.

We now prove the upper bound for (1.5.7). Writing

$$Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r\right) \leq Q_m\left(\max_{i=1}^m |\xi_i| \geq L\right) + Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r, \max_{i=1}^m |\xi_i| \leq L\right)$$

by Lemma 1.5.1, the upper bound follows easily provided we show that for all $L \geq x_1 \geq x_2 \cdots \geq x_r \geq 2$

$$\limsup_{m\to\infty} \frac{1}{m} \log Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r, \max_{i=1}^m |\xi_i| \leq L\right) \leq -I_r(x_1, x_2, \cdots, x_r) \tag{1.5.8}$$

To prove (1.5.8), let Q_{m-r}^m be a measure on $\mathbb{R}^{m-r}$ given by

$$Q_{m-r}^m(\xi \in \cdot) = Q_{m-r}\left((1 - \frac{r}{m})^{1/2}\xi \in \cdot\right).$$

We set

$$C_m^r = \left(1 - \frac{r}{m}\right)^{(m-r)^2/2} \frac{Z_{m-r}}{Z_m}$$

and for $x \in \mathbb{R}$ and $\mu \in \mathcal{P}(\mathbb{R})$ we define

$$\Phi(z, \mu) = 2\int \log|z - y|\mu(dy) - \frac{z^2}{2}$$

It was shown in [9] that $\Phi(z, \mu)$ is upper semi-continuous on $[-L, L] \times \mathcal{P}([-L, L])$ and continuous on $[x, y] \times \mathcal{P}([-L, L])$ for all $L, x, y \in \mathbb{R}$ such that $y > x > L > 2$. Using (1.5.1), we can write

$$\begin{aligned} &Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r, \max_{i=1}^m |\xi_i| \leq L\right) \\ &\leq C_m^r \frac{m!}{(m-r)!} \int_{x_1}^L \int_{x_2}^L \cdots \int_{x_r}^L \prod_{1\leq i<j\leq r} (\xi_i - \xi_j)^2 d\xi_r \cdots d\xi_2 d\xi_1 \\ &\times \int_{[-L,L]^{m-r}} e^{(m-r)\sum_{i=1}^r \Phi(\xi_i, \hat{\mu}_m^{m-r})} Q_{m-r}^m(d\xi_{r+1}, \cdots, d\xi_m) \end{aligned} \tag{1.5.9}$$

Here

$$\hat{\mu}_m^{m-r} = \frac{1}{m-r} \sum_{i=r+1}^m \delta_{\xi_i}$$

Let $B(\sigma,\delta)$ denote the open ball in $\mathcal{P}(\mathbb{R})$ of radius $\delta>0$ and center σ. We write $B_L(\sigma,\delta)=B(\sigma,\delta)\cap\mathcal{P}([-L,L])$. On the domain of the integration $(\xi_i-\xi_j)^2\le(2L)^2$ and $e^{(m-r)\Phi(\xi_i,\hat\mu_m^{m-r})}\le(2L)^{2(m-r)}$. Therefore (1.5.9) is bounded above by:

$$C_m^r\frac{m!}{(m-r)!}(2L)^{r(r-1)}\Bigg\{\prod_{i=1}^r\int_{x_i}^L e^{(m-r)\sup_{\mu\in B_L(\sigma,\delta)}\Phi(\xi_i,\mu)}d\xi_i \\ +(2L)^{r(2m-2r+1)}Q_{m-r}^m(\hat\mu_m^{m-r}\notin B(\sigma,\delta))\Bigg\} \tag{1.5.10}$$

To control the measure Q_{m-r}^m, observe that for all functions $h\in Lip_b(1)$ and m big enough:

$$\left|(m-r)^{-1}\sum_{i=r+1}^m\{h((1-rm^{-1})^{1/2}\xi_i)-h(\xi_i)\}\right|\le cm^{-1}\max_{i=r+1}^m|\xi_i|$$

for some $c\in\mathbb{R}^+$. It follows from Lemma 1 that $\hat\mu_m^{m-r}$ under Q_{m-r} and Q_{m-r}^m are exponentially equivalent as $m\to\infty$, so $\hat\mu_m^{m-r}$ under Q_{m-r}^m satisfies the same LDP as $\hat\mu_m^{m-r}$ under Q_{m-r}. Hence the second term of (1.5.10) is exponentially negligible for any $\delta>0$ and $L<\infty$. This implies that

$$\limsup_{m\to\infty}\frac{1}{m}\log Q_m\left(\xi_1^m\ge x_1,\cdots,\xi_r^m\ge x_r,\max_{i=1}^m|\xi_i|\le L\right) \\ \le\limsup_{m\to\infty}\frac{1}{m}\log C_m^r+\sum_{i=1}^r\lim_{\delta\downarrow0}\sup_{z\in[x_i,L],\mu\in B_L(\sigma,\delta)}\Phi(z,\mu) \tag{1.5.11}$$

Again from [9] the second term in (1.5.11) equals $-2\sum_{i=1}^r\int_2^{x_i}\sqrt{(z/2)^2-1}dz-r$. From the explicit expression of C_m^r it is easy to obtain
$\lim_{m\to\infty}m^{-1}\log C_m^r=r$. Combining these two claims we may complete the proof of the upper bound of (1.5.7).

To prove the lower bound, we fix $x_1\ge x_2\ge\cdots\ge x_r>t>2$. For $1\le i\le r$, fix

$y_i > x_i$, then for any $\delta > 0$

$$
\begin{aligned}
&Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r\right) \\
&\geq Q_m\left(\xi_1^m \in [x_1, y_1], \cdots, \xi_r^m \in [x_r, y_r], \max_{i=r+1}^{m} |\xi_i| \leq t\right) \\
&= C_m^r \int_{x_1}^{y_1} \int_{x_2}^{y_2} \cdots \int_{x_r}^{y_r} \prod_{i=1}^{r} e^{-r\xi_i^2/2} \prod_{1 \leq i < j \leq r} (\xi_i - \xi_j)^2 d\xi_r \cdots d\xi_2 d\xi_1 \\
&\times \int_{[-t,t]^{m-r}} e^{(m-r)\sum_{i=1}^{r} \Phi(\xi_i, \hat{\mu}_m^{m-r})} Q_{m-r}^m(d\xi_{r+1}, \cdots, d\xi_m) \\
&\geq K C_m^r \prod_{i=1}^{r} \exp\left\{(m-r) \inf_{z_i \in [x_i, y_i], \mu \in B_t(\sigma, \delta)} \Phi(z_i, \mu)\right\} Q_{m-r}^m(\hat{\mu}_m^{m-r} \in B_t(\sigma, \delta))
\end{aligned}
$$

for some $K = K(r, x_i, y_i) > 0$. Using the LDP for the measure $\hat{\mu}_m^{m-r}$ under Q_{m-r}^m we see that

$$
Q_{m-r}^m(\hat{\mu}_m^{m-r} \notin B(\sigma, \delta)) \to 0, \text{ as } m \to \infty
$$

By the symmetry of Q_{m-r}^m and by the upper bound of (1.5.7), we have

$$
Q_{m-r}^m(\hat{\mu}_m^{m-r} \notin \mathcal{P}((-t, t))) \leq 2Q_{m-r}(\xi_1^{m-r} \geq t) \to 0, \text{ as } m \to \infty
$$

Therefore, using the behavior of C_m^r again,

$$
\begin{aligned}
&\liminf_{m \to \infty} \frac{1}{m} \log Q_m\left(\xi_1^m \geq x_1, \cdots, \xi_r^m \geq x_r\right) \\
&\qquad \geq r + \sum_{i=1}^{r} \lim_{\delta \downarrow 0} \inf_{z_i \in [x_i, y_i], \mu \in B_t(\sigma, \delta)} \Phi(z_i, \mu) \qquad (1.5.12)
\end{aligned}
$$

letting $y_i \downarrow x_i$, we may get the lower bound of (1.5.7). ∎

From Theorem 1.5.1, and taking into account (1.5.4), we get a multidimensional LDP for the first r eigenvalues of the traceless GUE:

Theorem 1.5.2 *Let $r \in \mathbb{N}$, on $\mathcal{L}^r := \{(x_1, x_2, \cdots, x_r) \in \mathbb{R}^r : x_1 \geq x_2 \geq \cdots \geq x_r\}$, $(\xi_1^{m,0}, \xi_2^{m,0}, \cdots, \xi_r^{m,0})$ satisfies a LDP with speed m and a good rate function*

$$
I_r(x_1, x_2, \cdots, x_r) = \begin{cases} 2\sum_{i=1}^{r} \int_2^{x_i} \sqrt{(z/2)^2 - 1} dz, & \text{if } x_1 \geq x_2 \cdots \geq x_r \geq 2, \\ +\infty, & \text{otherwise.} \end{cases}
$$

Proof. From [9], $(\xi_1^m, \xi_2^m, \cdots, \xi_r^m)$ satisfies a LDP with speed m and rate function I_r on $\mathcal{L}^r$. To prove the validity of the same results for $(\xi_1^{m,0}, \xi_2^{m,0}, \cdots, \xi_r^{m,0})$, it is enough to show that

$$\limsup_{m\to\infty} \frac{1}{m} \log Q_m(\xi_r^{m,0} \le x) = -\infty, \tag{1.5.13}$$

for any $x < 2$, and since $I_r(x_1, x_2, \cdots, x_r)$ is continuous, increasing for any individual variable, on $\mathcal{L}^r \cap [2, \infty)^r$,

$$\lim_{m\to\infty} \frac{1}{m} \log Q_m\left(\xi_1^{m,0} \ge x_1, \cdots, \xi_r^{m,0} \ge x_r\right) = -2\sum_{i=1}^{r} \int_2^{x_i} \sqrt{(z/2)^2 - 1}dz, \tag{1.5.14}$$

for all $x_1 \ge x_2 \cdots \ge x_r \ge 2$.

First, for $x < 2$, let $\delta = 2 - x$, so

$$\begin{aligned} Q_m(\xi_r^{m,0} \le x) &\le Q_m(\xi_r^{m,0} + g_m \le x + \delta/2) + \mathbb{P}(g_m \ge \delta/2) \\ &= Q_m(\xi_r^m \le x + \delta/2) + \mathbb{P}(g_m \ge \delta/2). \end{aligned}$$

Since,

$$\mathbb{P}\left(g_m \ge \delta\right) \sim \frac{1}{\sqrt{2\pi}m\delta} e^{-m^2\delta^2/2}, \quad \text{as } m \to \infty, \tag{1.5.15}$$

(1.5.13) follows.

For (1.5.14), fix $x_1 \ge x_2 \cdots \ge x_r \ge 2$, for any $0 < \epsilon < x_r$, we have

$$\begin{aligned} &\limsup_{m\to\infty} \frac{1}{m} \log Q_m\left(\xi_1^{m,0} \ge x_1, \cdots, \xi_r^{m,0} \ge x_r\right) \\ &\le \limsup_{m\to\infty} \frac{1}{m} \log\left(Q_m\left(\xi_1^m \ge x_1 - \epsilon, \cdots, \xi_r^m \ge x_r - \epsilon\right) + \mathbb{P}\left(g_m \ge \epsilon\right)\right). \end{aligned}$$

Moreover,

$$Q_m\left(\xi_1^m \ge x_1 - \epsilon, \cdots, \xi_r^m \ge x_r - \epsilon\right) = \exp\{-m(I_r(x_1 - \epsilon, x_2 - \epsilon, \cdots, x_r - \epsilon) + o(1))\},$$

where $o(1)$ goes to 0 as m goes to infinity. So for fixed $0 < \epsilon < x_r$,

$$\frac{\mathbb{P}(g_m \ge \epsilon)}{Q_m\left(\xi_1^m \ge x_1 - \epsilon, \cdots, \xi_r^m \ge x_r - \epsilon\right)} \to 0, \quad m \to \infty,$$

hence,

$$\limsup_{m\to\infty}\frac{1}{m}\log Q_m\left(\xi_1^{m,0}\geq x_1,\cdots,\xi_r^{m,0}\geq x_r\right)\leq -I_r(x_1-\epsilon,x_2-\epsilon,\cdots,x_r-\epsilon).$$

Likewise,

$$\liminf_{m\to\infty}\frac{1}{m}\log Q_m\left(\xi_1^{m,0}\geq x_1,\cdots,\xi_r^{m,0}\geq x_r\right)\geq -I_r(x_1+\epsilon,x_2-\epsilon,\cdots,x_r+\epsilon).$$

Letting ϵ go to 0, the continuity of the rate function leads to (1.5.14). ■

For any $\mu\in\mathcal{P}(\mathbb{R})$, construct a discrete approximation of μ by setting

$$x_i^m=\inf\left\{x\in\mathbb{R}:\mu((-\infty,x])\geq\frac{i}{m+1}\right\},\quad 1\leq i\leq m, \tag{1.5.16}$$

and $\mu^m=\frac{1}{m}\sum_{i=1}^m\delta_{x_i^m}$ (note that the choice of the length $1/(m+1)$ of the intervals rather that $1/m$ is only made in order to insure that x_m^m is finite).

Using these discrete constructions, set:

$$\mathcal{X}=\left\{\mu\in\mathcal{P}(\mathbb{R}):\frac{1}{\sqrt{m}}\sum_{i=1}^m x_i^m\to 0,\ \text{as } m\to\infty\right\}, \tag{1.5.17}$$

and

$$\mathcal{P}_0(\mathbb{R})=\left\{\mu\in\mathcal{P}(\mathbb{R}):\int x d\mu(x)=0\right\}. \tag{1.5.18}$$

It is easy to see that $\mathcal{X}$ is a proper subset of $\mathcal{P}_0(\mathbb{R})$ since the condition in $\mathcal{X}$ implies that the mean of the measure is 0. With the help of this definition, following the proof in [10], we can get the large deviation principle for the spectral measure of the traceless GUE:

Theorem 1.5.3 *The spectral measure $\hat{\mu}_0^m=\frac{1}{m}\sum_{i=1}^m\delta_{\xi_i^0}$ satisfies a large deviation principle on $\mathcal{X}$ in the scale m^2 and with the good rate function I.*

Proof. Since this proof closely follows [10], it is just sketched here. Write the density of the eigenvalues as:

$$\begin{aligned}&Q_m(d\xi_1^0,d\xi_2^0,\cdots,d\xi_m^0)\\&=\frac{\sqrt{2\pi}}{Z_m}\exp\left\{-m^2\iint_{x\neq y}f(x,y)d\hat{\mu}_0^m(x)d\hat{\mu}_0^m(y)\right\}\prod_{i=1}^m e^{-\frac{{\xi_i^0}^2}{2}}d\xi_1^0\cdots\xi_{m-1}^0,\end{aligned}$$

where $\xi_m^0 = -\sum_{i=1}^{m-1} \xi_i^0$ and

$$f(x,y) = \frac{1}{4}(x^2+y^2) - \log|x-y|.$$

Let $\bar{Q}_m$ be the non-normalized positive measure $\bar{Q}_m = Z_m Q_m/\sqrt{2\pi}$. Via Stirling's formula,

$$\lim_{m\to\infty} \frac{1}{m^2} \log \frac{Z_m}{\sqrt{2\pi}} = \int_0^1 x \log x dx - \frac{1}{2} = -\frac{3}{4}, \tag{1.5.19}$$

so if under $\bar{Q}_m$, $\hat{\mu}_0^m$ satisfies a large deviation with rate function

$$J(\mu) = \iint f(x,y) d\mu(x) d\mu(y), \tag{1.5.20}$$

then combined with (1.5.19), this will lead to the statement of the theorem.

First, observe that for any Borel subset $A \subset \mathcal{X}$, any $N \in \mathbb{R}^+$,

$$\limsup_{m\to\infty} \frac{1}{m^2} \log\left(\bar{Q}_m(\hat{\mu}_0^m \in A)\right) \leq -\inf_{\mu\in A} \left\{ \iint f(x,y) \wedge N d\mu(x) d\mu(y) \right\}. \tag{1.5.21}$$

Moreover, from arguments as in [10], we get that $(\hat{\mu}_0^m)_{m\in\mathbb{N}}$ are exponentially tight under $\bar{Q}_m$ on $\mathcal{X}$. So we just need to prove $(\hat{\mu}_0^m)_{m\in\mathbb{N}}$ satisfies a weak large deviation principle with rate function $J(\mu)$ under the measure $\bar{Q}_m$. The upper bound is obvious, since $\mu \to \iint f(x,y) \wedge N d\mu(x) d\mu(y)$ is continuous for any $\mu \in \mathcal{X}$, so (1.5.21) shows that for any probability measure $\mu \in \mathcal{X}$,

$$\limsup_{\delta\to 0} \limsup_{m\to\infty} \frac{1}{m^2} \log\left(\bar{Q}_m(\hat{\mu}_0^m \in B(\mu,\delta))\right) \leq -\iint f(x,y) \wedge N d\mu(x) d\mu(y),$$

where $B(\mu,\delta)$ is an open ball of center μ and radius δ in $\mathcal{X}$, with the distance between two probability measures μ_1 and μ_2 in $\mathcal{X}$ is given by,

$$d(\mu_1,\mu_2) = \sup_{g\in Lip_b(1)} \left| \int g d\mu_1 - \int g d\mu_2 \right|,$$

and where for some fixed $b \geq 0$,

$$Lip_b(1) = \{g : \mathbb{R} \to \mathbb{R} : \|g\|_{Lip} \leq 1,\ \|g\|_\infty \leq b\}.$$

By monotone convergence,

$$\limsup_{\delta\to 0}\limsup_{m\to\infty}\frac{1}{m^2}\log\left(\bar{Q}_m(\hat{\mu}_0^m\in B(\mu,\delta))\right)\le -\iint f(x,y)d\mu(x)d\mu(y). \tag{1.5.22}$$

which finishes the proof of the upper bound.

To prove the lower bound, let $\nu\in\mathcal{X}$. Since $I(\nu)=+\infty$ if ν has an atom, we can assume without loss of generality here that it does not. Use the discrete construction (1.5.16) for ν with $\nu^m=\frac{1}{m}\sum_{i=1}^m\delta_{x_i^m}$. Since ν^m converges towards ν weakly with probability 1 as m goes to infinity, for any $\delta>0$ and m large enough, if we set $\Delta_m:=\{\xi_1^0\le\xi_2^0\le\cdots\le\xi_m^0\}$, then

$$\bar{Q}_m(\hat{\mu}_0^m\in B(\nu,\delta))\ge\bar{Q}_m\left(\left\{\max_{1\le i\le m-1}|\xi_i^0-x_i^m|<\frac{\delta}{2\sqrt{m}}\right\}\cap\Delta_m\right) \tag{1.5.23}$$

$$\begin{aligned}
&\ge\int_{\mathcal{T}(\xi_1,\cdots,\xi_m)}\exp\left\{-\frac{m}{2}\sum_{i=1}^m(\xi_i+x_i^m)^2\right\}\prod_{1\le i<j\le m}|\xi_i-\xi_j+x_i^m-x_j^m|^2\prod_{i=1}^{m-1}d\xi_i\\
&\ge\prod_{i+1<j}|x_i^m-x_j^m|^2\times\prod_{i=1}^{m-1}|x_{i+1}^m-x_i^m|\exp\left\{-\frac{m}{2}\sum_{i=1}^{m-1}(|x_i^m|+\frac{\delta}{\sqrt{m}})^2\right\}\\
&\times|x_m^m-x_{m-1}^m|\exp\left\{-m(\sum_{i=1}^{m-1}x_i^m)^2-m^2\delta^2\right\}\\
&\times\int_{\mathcal{T}(\xi_1,\cdots,\xi_m)}\prod_{i=1}^{m-2}|\xi_{i+1}-\xi_i|\prod_{i=1}^{m-1}d\xi_i,
\end{aligned} \tag{1.5.24}$$

where

$$\begin{aligned}
&\mathcal{T}(\xi_1,\cdots,\xi_m)\\
&:=\left\{\max_{1\le i\le m-1}|\xi_i|<\frac{\delta}{2\sqrt{m}},\ \xi_1\le\xi_2\le\cdots\le\xi_m,\ \sum_{i=1}^m\xi_i+\sum_{i=1}^m x_i^m=0\right\}.
\end{aligned}$$

The last term in the right hand side of (1.5.24) can be bounded from below by changing variables $\xi_1=x_1$ and $\xi_i-\xi_{i-1}=x_i,\ 2\le i\le m-1$. Set

$$\begin{aligned}
&\mathcal{R}(x_1,\cdots,x_{m-1})\\
&:=\left\{-\frac{\delta}{2\sqrt{m}}\le x_1\le-\frac{\delta}{4\sqrt{m}},\ \text{and } 0\le x_i\le-\frac{\delta}{4m^2}\ \text{for } 2\le i\le m-1\right\}.
\end{aligned}$$

Recalling that, $\frac{1}{\sqrt{m}}\sum_{i=1}^{m} x_i^m \to 0$,

$$\int_{\mathcal{T}(\xi_1,\cdots,\xi_m)} \prod_{i=1}^{m-2} |\xi_{i+1}-\xi_i| \prod_{i=1}^{m-1} d\xi_i \geq \int_{\mathcal{R}(x_1,\cdots,x_{m-1})} \prod_{i=2}^{m-1} |x_i| \prod_{i=1}^{m-1} dx_i$$
$$\geq \frac{\delta}{4\sqrt{m}} \left(\frac{1}{2}\left(\frac{\delta}{4m^2}\right)^2\right)^{m-2}. \tag{1.5.25}$$

Hence,

$$\bar{Q}_m(\hat{\mu}_0^m \in B(\nu,\delta)) \geq \prod_{i+1<j} |x_i^m - x_j^m|^2 \prod_{i=1}^{m-1} |x_{i+1}^m - x_i^m| \exp\left\{-\frac{m}{2}\sum_{i=1}^{m}(x_i^m)^2\right\}$$
$$\times |x_m^m - x_{m-1}^m| \frac{\delta}{4\sqrt{m}} \left(\frac{1}{2}\left(\frac{\delta}{4m^2}\right)^2\right)^{m-2} \exp\left\{-\sqrt{m}\delta\sum_{i=1}^{m}|x_i^m| - \delta^2\right\}. \tag{1.5.26}$$

Now by same arguments as in [10], we get

$$\liminf_{\delta\to 0}\liminf_{m\to\infty} \frac{1}{m^2}\log\left(\bar{Q}_m(\hat{\mu}_0^m \in B(\nu,\delta))\right) \geq -\iint f(x,y)d\nu(x)d\nu(y). \tag{1.5.27}$$

Combining (1.5.22) and (1.5.27), the weak large deviation principle is proved, finishing the whole proof. ∎

We are now ready to give the large deviation for $\xi_1^{m,0}$ when it is on the left of its mean. Let $\mathcal{M}((-\infty,x])$ be the set of all probability measures on $(-\infty,x], x\in\mathbb{R}$, let $\mathcal{M}_{\mathcal{X}}((-\infty,x]) = \mathcal{M}((-\infty,x])\cap\mathcal{X}$, and let $\mathcal{M}_0((-\infty,x]) = \mathcal{M}((-\infty,x])\cap\mathcal{P}_0(\mathbb{R})$. Since $\{\xi_1^{m,0}\leq x\} = \{\hat{\mu}_0^m \in \mathcal{M}_{\mathcal{X}}((-\infty,x])\}$, then for any $x\leq 2$,

$$\lim_{m\to\infty}\frac{1}{m^2}\log\mathbb{P}\left(\xi_1^{m,0}\leq x\right) = -\inf_{\mu\in\mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu). \tag{1.5.28}$$

For each $x\in\mathbb{R}$, let

$$K(x) = \inf_{\mu\in\mathcal{M}_0((-\infty,x])} I(\mu). \tag{1.5.29}$$

When $x\geq 2$, the semicircular law σ is both in $\mathcal{M}_{\mathcal{X}}((-\infty,x])$ and $\mathcal{M}_0((-\infty,x])$, and so $\inf_{\mu\in\mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu) = K(x) = I(\sigma) = 0$. Moreover, when $x\leq 0$, and since both $\mathcal{M}_{\mathcal{X}}((-\infty,x])$ and $\mathcal{M}_0((-\infty,x])$ are empty, it follows that $\inf_{\mu\in\mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu) = K(x) = I(\sigma) = +\infty$.

When $0 < x \leq 2$, and from arguments as in [23], it is next shown that K is continuous. Indeed, for any $y < 0$ and $0 < x \leq 2$, let

$$J_\mu(y,x) = \frac{1}{2}\int_y^x u^2 d\mu(u) - \int_y^x \int_y^x \log|u-t| d\mu(u) d\mu(t) - \frac{3}{4}, \tag{1.5.30}$$

and let ν_x be the minimizer of $I(\mu)$ on $\mathcal{M}_0((-\infty, x])$, then for any $0 < \epsilon < x$, we have

$$K(x) \leq K(x-\epsilon) \leq \frac{J_{\nu_x}(y_\epsilon, x-\epsilon)}{\nu_x^2([y_\epsilon, x-\epsilon])}, \tag{1.5.31}$$

where y_ϵ is the value which satisfies

$$\int_{y_\epsilon}^{x-\epsilon} t d\nu_x(t) = 0.$$

Since the right hand side of (1.5.31) converges to $K(x)$, as ϵ converges to 0, the left continuity of K is proved.

To show the right continuity, notice that by a simple change of variables,

$$K(x) = \inf_{\mu \in \mathcal{M}_0((-\infty, x+\epsilon])} J_\mu^\epsilon(x),$$

where

$$J_\mu^\epsilon(x) = \frac{1}{2}\int_{-\infty}^{x+\epsilon} (u-\epsilon)^2 d\mu(u) - \int_{-\infty}^{x+\epsilon}\int_{-\infty}^{x+\epsilon} \log|u-t| d\mu(u) d\mu(t) - \frac{3}{4}.$$

Therefore,

$$0 \leq K(x) - K(x+\epsilon) \leq J_{\nu_{x+\epsilon}}^\epsilon(x) - K(x+\epsilon) = \frac{1}{2}\epsilon^2,$$

thus by letting ϵ go to 0, the right continuity of K follows. Likewise, it can be proved that $\inf_{\mu \in \mathcal{M}_\mathcal{X}((-\infty, x])} I(\mu)$ is right continuous with respect to x.

Next we need a lemma which, when combined with (1.5.28), gives

$$\lim_{m\to\infty} \frac{1}{m^2} \log \mathbb{P}\left(\xi_1^{m,0} \leq x\right) = -K(x), \tag{1.5.32}$$

for any $x \leq 2$.

Our next lemma and its proof benefited from Ionel Popescu input.

Lemma 1.5.2 *For any $x \in \mathbb{R}$,*

$$\inf_{\mu \in \mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu) = K(x). \tag{1.5.33}$$

Proof. For $x \geq 2$, both sides in (1.5.33) are equal to zero, we thus just need to consider the case $x < 2$. First, since $\mathcal{X}$ is a proper subset of $\mathcal{P}_0(\mathbb{R})$,

$$K(x) \leq \inf_{\mu \in \mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu). \tag{1.5.34}$$

Next, we need to prove

$$K(x) \geq \inf_{\mu \in \mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu). \tag{1.5.35}$$

From Theorem 1.10 and Theorem 1.11 of Chapter IV in [35], we know that there is a unique probability measure, call it μ_0, which minimizes $I(\mu)$ for all $\mu \in \mathcal{M}_0((-\infty, x])$, and the support of μ_0 is an interval, denoted as $[a, b]$ (with $b \leq x$). Since μ_0 is atomless, its distribution function F is continuous, increasing with $F(a) = 0$ and $F(b) = 1$. Moreover, since μ_0 has zero mean, $\int_0^1 F^{-1}(x)dx = 0$, where F^{-1}, the inverse of F, is continuous and increasing on $[0, 1]$, with $F^{-1}(0) = a$ and $F^{-1}(1) = b$.

Now for any integer $n \geq 2$, construct an approximation to F^{-1} as follows: for $i/n \leq x \leq (i + 1)/n$, let

$$G_n^+(x) = \begin{cases} n\left(F^{-1}(\frac{i+2}{n})\left(x - \frac{i}{n}\right) + F^{-1}(\frac{i+1}{n})\left(\frac{i+1}{n} - x\right)\right), & \text{if } 0 \leq i \leq n-2, \\ b + x - \frac{i}{n}, & \text{if } i = n-1, \end{cases}$$

and let

$$G_n^-(x) = \begin{cases} n\left(F^{-1}(\frac{i}{n})\left(x - \frac{i}{n}\right) + F^{-1}(\frac{i-1}{n})\left(\frac{i+1}{n} - x\right)\right), & \text{if } 1 \leq i \leq n-1, \\ a + x - \frac{i+1}{n}, & \text{if } i = 0. \end{cases}$$

From this construction, $\int_0^1 G_n^+(x)dx > 0$ and $\int_0^1 G_n^-(x)dx < 0$. Next, let

$$\gamma_n^+ = \frac{-\int_0^1 G_n^-(x)dx}{\int_0^1 G_n^+(x)dx - \int_0^1 G_n^-(x)dx}, \ \gamma_n^- = \frac{\int_0^1 G_n^+(x)dx}{\int_0^1 G_n^+(x)dx - \int_0^1 G_n^-(x)dx},$$

and let

$$G_n(x) = \gamma_n^+ G_n^+(x) + \gamma_n^- G_n^-(x).$$

Then,

$$\int_0^1 G_n(x)dx = 0,$$

and since G_n is piecewisely linear, it is Lipschitz continuous. Let μ_n be the probability measure whose distribution function is the inverse function of G_n, the Lipschitz continuity of G_n yields that $\mu_n \in \mathcal{X}$, for any $n \geq 2$. From its construction, we know that μ_n is supported on $[a - 1/n, b + 1/n]$, and μ_n converges to μ_0 weakly as n goes to infinity, thus

$$\lim_{n\to\infty} \int x^2 d\mu_n(x) = \int x^2 d\mu_0(x). \tag{1.5.36}$$

For the second term on the right side of (1.5.5),

$$\iint \log|x-y| d\mu(x)d\mu(y) = 2\iint_{x<y} \log(y-x) d\mu(x)d\mu(y), \tag{1.5.37}$$

let

$$\frac{1}{n^2}\sum_{i<j} \log\left(F^{-1}\left(\frac{j+1}{n}\right) - F^{-1}\left(\frac{i}{n}\right)\right) + \frac{1}{2n^2}\sum_{i=0}^{n-1} \log\left(F^{-1}\left(\frac{i+1}{n}\right) - F^{-1}\left(\frac{i}{n}\right)\right) \tag{1.5.38}$$

and

$$\frac{1}{n^2}\sum_{i<j} \log\left(G_n\left(\frac{j+1}{n}\right) - G_n\left(\frac{i}{n}\right)\right) + \frac{1}{2n^2}\sum_{i=0}^{n-1} \log\left(G_n\left(\frac{i+1}{n}\right) - G_n\left(\frac{i}{n}\right)\right) \tag{1.5.39}$$

be Riemann sum approximations of $\iint_{x<y} \log(y-x)d\mu_0(x)d\mu_0(y)$ and $\iint_{x<y} \log(y-x)d\mu_n(x)d\mu_n(y)$ respectively. For any $i \leq j$,

$$\begin{aligned}
&\log\left(G_n\left(\frac{j+1}{n}\right) - G_n\left(\frac{i}{n}\right)\right) \\
&\geq \gamma_n^+ \log\left(G_n^+\left(\frac{j+1}{n}\right) - G_n^+\left(\frac{i}{n}\right)\right) + \gamma_n^- \log\left(G_n^-\left(\frac{j+1}{n}\right) - G_n^-\left(\frac{i}{n}\right)\right),
\end{aligned} \tag{1.5.40}$$

and moreover, for any $1 \leq i \leq j \leq n-2$,

$$\begin{aligned}&\log\left(G_n\left(\frac{j+1}{n}\right)-G_n\left(\frac{i}{n}\right)\right)\\&\geq \gamma_n^+\log\left(F^{-1}\left(\frac{j+2}{n}\right)-F^{-1}\left(\frac{i+1}{n}\right)\right)+\gamma_n^-\log\left(F^{-1}\left(\frac{j}{n}\right)-F^{-1}\left(\frac{i-1}{n}\right)\right).\end{aligned} \tag{1.5.41}$$

If $\iint_{x<y}\log(y-x)d\mu_0(x)d\mu_0(y)=-\infty$, (1.5.35) is trivially true, so assume this integral is finite. Moreover, since $\gamma_n^+ + \gamma_n^- = 1$,

$$\liminf_{n\to\infty}\left(-\iint\log|x-y|d\mu_n(x)d\mu_n(y)\right)\leq -\iint\log|x-y|d\mu_0(x)d\mu_0(y), \tag{1.5.42}$$

and combining (1.5.36) and (1.5.42), gives

$$\liminf_{n\to\infty} I(\mu_n)\leq I(\mu_0).$$

Since μ_n is supported on $[a-1/n, b+1/n]$ and from the right continuity with respect to x of $\inf_{\mu\in\mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu)$, we know that

$$K(x)\geq \inf_{\mu\in\mathcal{M}_{\mathcal{X}}((-\infty,x])} I(\mu),$$

which finishes the proof. ∎

To finish, we obtain the large deviations for the first r eigenvalues of the traceless GUE when at least one of them is on the left of the asymptotic mean:

Corollary 1.5.1 *For $x_r \leq x_{r-1} \leq \cdots \leq x_1$, and $x_r \leq 2$, we have*

$$\lim_{m\to\infty}\frac{1}{m^2}\log\mathbb{P}\left(\xi_1^{m,0}\leq x_1,\cdots,\xi_r^{m,0}\leq x_r\right)=-K(x_r),$$

Proof. Since $(\xi_1^{m,0},\xi_2^{m,0},\cdots,\xi_m^{m,0})$ is the nonincreasing rearrangement of $(\xi_1^0,\xi_2^0,\cdots,\xi_m^0)$, set

$$\begin{aligned}A&:=\mathbb{P}\left(\xi_1^{m,0}\leq x_1,\cdots,\xi_r^{m,0}\leq x_r\right),\\B&:=\mathbb{P}\left(\xi_1^0\leq x_1,\cdots,\xi_r^0\leq x_r,\xi_{r+1}^0\leq x_r,\cdots,\xi_m^0\leq x_r\right),\end{aligned}$$

then

$$B \leq A \leq \frac{m!}{(m-r+1)!(r-1)!} B \leq m^r B. \tag{1.5.43}$$

Therefore,

$$\lim_{m\to\infty} \frac{1}{m^2} \log A = \lim_{m\to\infty} \frac{1}{m^2} \log B. \tag{1.5.44}$$

Changing variables:

$$\xi_i^0 - (x_i - x_r) = \eta_i, \quad \text{for} \;\; 1 \leq i \leq r-1,$$

$$\xi_i^0 = \eta_i, \quad \text{for} \;\; r \leq i \leq m,$$

we then have:

$$B = \mathbb{P}\left(\eta_i \leq x_r, \; 1 \leq i \leq m\right).$$

Considering the two measures $\frac{1}{m}\sum_{i=1}^m \xi_i^0$ and $\frac{1}{m}\sum_{i=1}^m \eta_i$, for any bounded and Lipschitz function g, we have

$$\frac{1}{m}\left|\sum_{i=1}^m g(\xi_i^0) - \sum_{i=1}^m g(\eta_i)\right| \leq \frac{1}{m}\sum_{i=1}^m |\xi_i^0 - \eta_i| \to 0, \quad \text{as} \;\; m \to \infty$$

so $\frac{1}{m}\sum_{i=1}^m \xi_i^0$ and $\frac{1}{m}\sum_{i=1}^m \eta_i$ are exponentially equivalent, and Theorem 1.5.3 also applies for the latter (see Theorem 4.2.13 in [16]). So from (1.5.32), we have

$$\lim_{m\to\infty} \frac{1}{m^2} \log B = -K(x_r),$$

and (1.5.44) finishes the proof. ∎

CHAPTER II

ON THE ORDER OF THE CENTRAL MOMENTS OF THE LENGTH OF THE LONGEST COMMON SUBSEQUENCE

2.1 Introduction and results

Let $X = (X_i)_{i\geq 1}$ and $Y = (Y_i)_{i\geq 1}$ be two independent sequences of iid random variables taking values in a finite alphabet $\mathcal{A}_m = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$, with $\mathbb{P}(X_1 = \alpha_k) = \mathbb{P}(Y_1 = \alpha_k) = p_k$, $k = 1, 2, \cdots, m$. Let LC_n be the length of the longest common subsequence of $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, *i.e.*, LC_n is the largest k such that there exist $1 \leq i_1 < i_2 < \cdots < i_k \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_k \leq n$, such that $X_{i_1} = Y_{j_1}, X_{i_2} = Y_{j_2}, \cdots, X_{i_k} = Y_{j_k}$.

The study of the asymptotic behavior of LC_n has a long history starting with the well known result of Chvátal and Sankoff [14] which asserts that

$$\lim_{n\to\infty} \frac{\mathbb{E}LC_n}{n} = \gamma^*. \tag{2.1.1}$$

However, the exact value of γ^*, which depends on the distribution of X_1 and on the size of the alphabet is still unknown, even in "simple cases" such as for uniform Bernoulli random variables. This first asymptotic result was sharpened by Alexander in [1] and [2] where the speed of convergence to γ^* in (2.1.1) is investigated, and where it is shown that,

$$\gamma^* n - C\sqrt{n \log n} \leq \mathbb{E}LC_n \leq \gamma^* n, \tag{2.1.2}$$

where $C > 0$ is a constant depending neither on n nor on the distribution of X_1. Next, as far as the order of the variance is concerned, Steele [37] first proved that

$VarLC_n \leq n$, but finding the order of the lower bound is more illusive. In various instances, where there is some "bias" such as for an asymmetric scoring function or highly asymmetric Bernoulli random variables, the lower bound is shown to be of order n ([13], [21], [30]). This is also the case if the sequences are iid uniform and contain sparse long blocks, which is in some sense a situation as close as we want to the iid uniform one (see [6]).

We investigate below the r-th central moment of LC_n, when all but one of the letters are drawn with small probabilities, and prove:

Theorem 2.1.1 *Let $1 \leq r < +\infty$, and let $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ be two independent sequences of iid random variables with values in $\mathcal{A}_m = \{\alpha_1, \alpha_2, \cdots, \alpha_m\}$, and with $\mathbb{P}(X_1 = \alpha_k) = p_k$, $k = 1, 2, \cdots, m$. Let $p_{j_0} > 1/2$, for some $j_0 \in \{1, \cdots, m\}$ and let $\max_{j\neq j_0} p_j \leq K/m$, where $K = 2^{-6}10^{-1}e^{-21}$. Then there exists a constant $C > 0$ depending on p_{j_0} and $\max_{j\neq j_0} p_j$, but independent of $n \in \mathbb{N}$, such that,*

$$\mathbb{M}_r(LC_n) := \mathbb{E}\,|LC_n - \mathbb{E}LC_n|^r \geq Cn^{\frac{r}{2}}. \tag{2.1.3}$$

The above result provides yet another instance where the variance is linear in the length of the sequences since, using the upper bound of [37], the lower bound (2.1.3) can clearly be complemented by

$$\mathbb{M}_r(LC_n) \leq (VarLC_n)^{\frac{r}{2}} \leq n^{\frac{r}{2}}, \tag{2.1.4}$$

$0 < r \leq 2$.

As far as the content of the paper is concerned, in Section 2.2 we give a proof of Theorem 2.1.1, which relies on the key preliminary result Theorem 2.2.1, whose proof is given in Section 2.3.

2.2 Proof of Theorem 2.1.1

Throughout the paper, by finite sequences X and Y of length n, and when this does not cause confusion, it is meant that $X = (X_i)_{1\leq i\leq n}$ and $Y = (Y_i)_{1\leq i\leq n}$.

The strategy of proof to obtain our lower bound is to first represent LC_n as a random function of the number of α_{j_0} in the sequences. This random function satisfies a local reversed Lipschitz condition, as n goes to infinity, which ultimately gives the lower bound in Theorem 2.1.1. To do so, pick a letter equiprobably at random from all the non-α_{j_0} letters in either one of the two finite sequences, of length n, X and Y, then change it to the most likely letter α_{j_0}, and call the two new finite sequences $\tilde{X}$ and $\tilde{Y}$. Then, the length of the longest common subsequence of $\tilde{X}$ and $\tilde{Y}$, denoted by $\widetilde{LC_n}$, tends, on an event of high probability, to be larger than LC_n. This fact is given in the following theorem.

Theorem 2.2.1 *Let the hypothesis of Theorem 2.1.1 hold. Then, there exists a set $\mathcal{B}_n \subset \mathcal{A}_m^n \times \mathcal{A}_m^n$, such that for all $n \geq 1$,*

$$\mathbb{P}\left((X,Y) \in \mathcal{B}_n\right) \geq 1 - 204 \exp\left(-\frac{n(\max_{j \neq j_0} p_j)^6}{5}\right), \tag{2.2.1}$$

and such that for all $(x,y) \in \mathcal{B}_n$,

$$\mathbb{P}(\widetilde{LC_n} - LC_n = 1 | X = x, Y = y) \geq K_2, \tag{2.2.2}$$

$$\mathbb{P}(\widetilde{LC_n} - LC_n = -1 | X = x, Y = y) \leq \frac{K_2}{2}, \tag{2.2.3}$$

where $K_2 = K_1/m$, with $K_1 = 2^{-4}10^{-1}e^{-16}$.

The proof of Theorem 2.2.1 is given in the next section, but we indicate next how it leads to the lower bound on $\mathbb{M}_r(LC_n)$ obtained in Theorem 2.1.1.

From now on, assume without loss of generality that $p_1 > 1/2$ and that $p_2 = \max_{2 \leq j \leq m} p_j$.

We start with a few definitions. For the two finite random sequences $X = (X_i)_{1 \leq i \leq n}$ and $Y = (Y_i)_{1 \leq i \leq n}$, let N_1 be the total number of letters α_1 present in them. By induction, define next a finite collection of pairs of random sequences $(X^k, Y^k)_{0 \leq k \leq 2n}$ as follows: First, let $X^0 = (X_i^0)_{1 \leq i \leq n}$ and $Y^0 = (Y_i^0)_{1 \leq i \leq n}$ be independent, with X_i^0 and Y_i^0, $i = 1, \cdots, n$, iid random variables with values in $\{\alpha_2, \cdots, \alpha_m\}$

and such that $\mathbb{P}(X_1^0 = \alpha_k) = \mathbb{P}(Y_1^0 = \alpha_k) = p_k/(1-p_1)$, $2 \le k \le m$. In other words, X^0, Y^0 are two independent finite sequences of iid random variables with common law $\mathcal{L}(X, Y|N_1 = 0)$. Once (X^k, Y^k) is defined, let (X^{k+1}, Y^{k+1}) be the pair of finite random sequences obtained by taking one letter with equal probability from all the letters $\alpha_2, \alpha_3, \cdots, \alpha_m$ in the pair (X^k, Y^k) and replacing it with α_1. Then, let $LC_n(k)$ denote the length of the longest common subsequence of X^k and Y^k. Our first lemma shows that the law of (X^k, Y^k) is the same as the law of (X, Y) conditional on $N_1 = k$, and therefore the law of $LC_n(k)$ is the same as the conditional law of LC_n given $N_1 = k$.

Lemma 2.2.1 *For $k = 0, 1, \cdots, 2n$,*

$$\mathcal{L}(X^k, Y^k) = \mathcal{L}(X, Y|N_1 = k). \tag{2.2.4}$$

Proof. The proof is by induction on k. By definition, (X^0, Y^0) has the same law as (X, Y) conditional on $N_1 = 0$. Now assume (2.2.4) is true for k, *i.e.*, for any $(\alpha_{j_1}, \cdots, \alpha_{j_{2n}}) \in \mathcal{A}_m^n \times \mathcal{A}_m^n$, with $J_1 = k$,

$$\mathbb{P}\left((X_1^k, \cdots, X_n^k, Y_1^k, \cdots, Y_n^k) = (\alpha_{j_1}, \cdots, \alpha_{j_{2n}})\right) = \binom{n}{k}^{-1} \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{J_\ell}, \tag{2.2.5}$$

where $J_\ell = |\{1 \le i \le 2n : \ \alpha_{j_i} = \alpha_\ell\}|$, for any $1 \le \ell \le m$. Then, for $k+1$, for any $(\alpha_{j_1}, \cdots, \alpha_{j_{2n}}) \in \mathcal{A}_m^n \times \mathcal{A}_m^n$, with $J_1 = k+1$,

$$\mathbb{P}\left((X_1^{k+1}, \cdots, X_n^{k+1}, Y_1^{k+1}, \cdots, Y_n^{k+1}) = (\alpha_{j_1}, \cdots, \alpha_{j_{2n}})\right) =$$
$$\sum_{i=1}^{k+1} \mathbb{P}\left((X_1^{k+1}, \cdots, X_n^{k+1}, Y_1^{k+1}, \cdots, Y_n^{k+1}) = (\alpha_{j_1}, \cdots, \alpha_{j_{2n}})|B_i^{k+1}\right) \mathbb{P}(B_i^{k+1}), \tag{2.2.6}$$

where B_i^{k+1} is the event that the i-th α_1 in $(\alpha_{j_1}, \cdots, \alpha_{j_{2n}})$ is changed from a non-α_1 letter when passing from (X^k, Y^k) to (X^{k+1}, Y^{k+1}), for $1 \le i \le k+1$. Conditional on B_i^{k+1}, the i-th α_1 in $(\alpha_{j_1}, \cdots, \alpha_{j_{2n}})$ can be changed from any letter in $\{\alpha_2, \alpha_3, \cdots, \alpha_m\}$, for each corresponding instance, the probability of the corresponding (X^k, Y^k) is

$$\binom{n}{k}^{-1} \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{J_\ell} \left(\frac{p_\jmath}{1-p_1}\right),$$

$2 \leq \jmath \leq m$. Thus,

$$\begin{aligned}
&\mathbb{P}\left((X_1^{k+1}, \cdots, X_n^{k+1}, Y_1^{k+1}, \cdots, Y_n^{k+1}) = (\alpha_{j_1}, \cdots, \alpha_{j_{2n}}) | B_i^{k+1}\right) \mathbb{P}(B_i^{k+1}) \\
&= \binom{n}{k}^{-1} \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{J_\ell} \left(\sum_{\jmath=2}^{m} \frac{p_\jmath}{1-p_1}\right) \frac{1}{n-k},
\end{aligned}$$

plugging this into (2.2.6), gives

$$\begin{aligned}
&\mathbb{P}\left((X_1^{k+1}, \cdots, X_n^{k+1}, Y_1^{k+1}, \cdots, Y_n^{k+1}) = (\alpha_{j_1}, \cdots, \alpha_{j_{2n}})\right) \\
&\qquad = \binom{n}{k+1}^{-1} \prod_{\ell=2}^{m} \left(\frac{p_\ell}{1-p_1}\right)^{J_\ell}, \quad (2.2.7)
\end{aligned}$$

and this finishes the proof. ∎

Since the pairs $\{(X^k, Y^k)\}_{0 \leq k \leq 2n}$ are independent of the random variable N_1, (X^{N_1}, Y^{N_1}) has the same law as (X, Y). Therefore, $LC_n(N_1)$, the length of the longest common subsequence of (X^{N_1}, Y^{N_1}), has the same law as LC_n, and thus,

$$\mathbb{M}_r(LC_n(N_1)) = \mathbb{M}_r(LC_n). \quad (2.2.8)$$

To prove Theorem 2.1.1, we also need the following simple inequality valid for locally reversed Lipschitz functions.

Lemma 2.2.2 *Let $f : D \to \mathbb{Z}$ satisfies a local reversed Lipschitz condition, i.e., let f be such that for any $i, j \in D$ with $j \geq i + \ell$,*

$$f(j) - f(i) \geq c(j - i),$$

where c and ℓ are two positive constants, and let T be a D-valued random variable such that $\mathbb{E}|f(T)| < +\infty$, then for any $r \geq 1$,

$$\mathbb{M}_r(f(T)) \geq \left(\frac{c}{2}\right)^r \left(\mathbb{M}_r(T) - \ell^r\right). \quad (2.2.9)$$

Proof. By convexity, if $\widehat{T}$ is an independent copy of T, then for any $r \geq 1$,

$$\mathbb{M}_r(T) \leq \mathbb{E}(|T - \widehat{T}|^r) \leq 2^r \mathbb{M}_r(T), \quad (2.2.10)$$

which in turn implies,

$$\begin{aligned}\mathbb{M}_r(f(T)) &\geq \frac{1}{2^r}\mathbb{E}(|f(T)-f(\widehat{T})|^r)\\ &\geq \left(\frac{c}{2}\right)^r\left(\mathbb{E}((T-\widehat{T})^r\mathbf{1}_{T-\widehat{T}\geq\ell})+\mathbb{E}((\widehat{T}-T)^r\mathbf{1}_{\widehat{T}-T\geq\ell})\right)\\ &\geq \left(\frac{c}{2}\right)^r\left(\mathbb{E}(|T-\widehat{T}|^r)-\ell^r\right)\\ &\geq \left(\frac{c}{2}\right)^r\left(\mathbb{M}_r(T)-\ell^r\right).\end{aligned}$$

■

For any random variable U and random vector V with finite r-th moment, $r \geq 1$, let

$$\mathbb{M}_r(U|V) := \mathbb{E}\left(|U-\mathbb{E}\left(U|V\right)|^r\middle|V\right),$$

then,

$$\begin{aligned}\mathbb{E}\left(\mathbb{M}_r(U|V)\right) &= \mathbb{E}\left(|U-\mathbb{E}(U)+\mathbb{E}(U)-\mathbb{E}(U|V)|^r\right)\\ &\leq 2^{r-1}\left(\mathbb{E}\left(|U-\mathbb{E}(U)|^r\right)+\mathbb{E}\left(|\mathbb{E}(U|V)-\mathbb{E}(U)|^r\right)\right)\\ &= 2^{r-1}\left(\mathbb{M}_r(U)+\mathbb{E}\left(|\mathbb{E}\left(U-\mathbb{E}(U)|V\right)|^r\right)\right)\\ &\leq 2^r\mathbb{M}_r(U),\end{aligned}$$

hence

$$\mathbb{M}_r(U) \geq \frac{1}{2^r}\mathbb{E}(\mathbb{M}_r(U|V)), \tag{2.2.11}$$

which in our setting implies that, for each $n \geq 1$,

$$\mathbb{M}_r(LC_n(N_1)) \geq \frac{1}{2^r}\mathbb{E}(\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n})). \tag{2.2.12}$$

But, N_1 is independent of $(LC_n(k))_{0\leq k\leq 2n}$, and so from (2.2.11),

$$\begin{aligned}&\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n})\\ &\qquad\geq \frac{1}{2^r}\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n}, N_1\in I)\mathbb{P}(N_1\in I),\end{aligned} \tag{2.2.13}$$

where I is the interval

$$I = \left[2np_1 - \sqrt{2n(1-p_1)p_1}, 2np_1 + \sqrt{2n(1-p_1)p_1}\right]. \tag{2.2.14}$$

.

Likewise,

$$\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n}, N_1 \in I) \\ \geq \frac{1}{2^r}\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n}, N_1 \in I \cap O_n)\mathbb{P}(O_n), \tag{2.2.15}$$

where for $n \geq 1$, O_n is the event that for all $i, j \in I$, such that $j + \ell(n) \leq i$,

$$LC_n(i) - LC_n(j) \geq \frac{K_2}{4}|i-j|, \tag{2.2.16}$$

where K_2 is given in Theorem 2.2.1 and where $\ell(n)$ is to be chosen later. In other words, on the event O_n the random function $LC_n(\cdot)$ has a slope of at least $K_2/4$ on the interval I, when i and j are at least $\ell(n)$ away from each other.

Combining (2.2.12), (2.2.13) and (2.2.15) leads to

$$\mathbb{M}_r(LC_n(N_1)) \\ \geq \frac{1}{8^r}\mathbb{E}(\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n}, N_1 \in I \cap O_n))\mathbb{P}(N_1 \in I)\mathbb{P}(O_n), \tag{2.2.17}$$

and it remains to estimate the three terms on the right hand side of (2.2.17). For the first one, from the very definition of the event O_n, applying Lemma 2.2.2, and since N_1 is independent of $(LC_n(k))_{0\leq k\leq 2n}$,

$$\mathbb{E}(\mathbb{M}_r(LC_n(N_1)|(LC_n(k))_{0\leq k\leq 2n}, N_1 \in I \cap O_n)) \\ \geq \left(\frac{K_2}{8}\right)^r \left(\mathbb{M}_r(N_1|N_1 \in I) - \ell(n)^r\right). \tag{2.2.18}$$

Next, from the Berry-Esséen inequality, for all $n \geq 1$,

$$\left|\mathbb{P}(N_1 \in I) - \frac{1}{\sqrt{2\pi}}\int_{-1}^{1} e^{-\frac{x^2}{2}}dx\right| \leq \frac{1}{\sqrt{2np_1(1-p_1)}}, \tag{2.2.19}$$

and thus

$$\mathbb{M}_r(N_1|N_1 \in I) \geq (2np_1(1-p_1))^{\frac{r}{2}} \frac{\int_{-1}^{1} |x|^r e^{-\frac{x^2}{2}} dx - 2\sqrt{\pi}/\sqrt{np_1(1-p_1)}}{\int_{-1}^{1} e^{-\frac{x^2}{2}} dx + \sqrt{\pi}/\sqrt{np_1(1-p_1)}}. \quad (2.2.20)$$

Choosing $33 \log n/K_2^2 \leq \ell(n) = o(\sqrt{n})$, combining (2.2.17)-(2.2.20) and the estimate on $\mathbb{P}(O_n)$, proved in the next lemma, will finish the proof of the lower bound in (2.1.3) provided that $\ell(n) \geq 33 \log n/K_2^2$ and that Theorem 2.2.1 holds true.

Lemma 2.2.3 *Let $p_2 \leq K/m$, with $K = 2^{-6}10^{-1}e^{-21}$, then for all $n \geq 1$,*

$$\mathbb{P}(O_n) \geq 1 - \left(816\sqrt{\pi}e^2 n \exp\left(-\frac{np_2^6}{5}\right) + 2n \exp\left(-\frac{K_2^2 \ell(n)}{32}\right)\right). \quad (2.2.21)$$

Proof. Let $A_n := \{(X,Y) \in \mathcal{B}_n\}$ and let $A_n^k := \{(X^k, Y^k) \in \mathcal{B}_n\}$. Then,

$$\mathbb{P}\left(\left(\bigcap_{k\in I} A_n^k\right)^c\right) \leq \sum_{k\in I} \mathbb{P}\left(\left(A_n^k\right)^c\right) = \sum_{k\in I} \mathbb{P}\left(A_n^c | N_1 = k\right) \leq \sum_{k\in I} \frac{\mathbb{P}(A_n^c)}{\mathbb{P}(N_1 = k)}. \quad (2.2.22)$$

By stirling's formula, for all $k \in I$ and $n \geq 1$,

$$\begin{aligned}\mathbb{P}(N_1 = k) &= \binom{2n}{k} p_1^k (1-p_1)^{2n-k} \\ &\geq \frac{1}{\sqrt{2\pi}e^2} \frac{(2n)^{2n+1/2}}{k^{k+1/2}(2n-k)^{2n-k+1/2}} p_1^k (1-p_1)^{2n-k}.\end{aligned}$$

Let

$$\eta(k,n,p_1) := \frac{1}{\sqrt{2\pi}e^2} \frac{(2n)^{2n+1/2}}{k^{k+1/2}(2n-k)^{2n-k+1/2}} p_1^k (1-p_1)^{2n-k},$$

then for all $k \in I$ and $p_1 \geq 3/4$ (which holds true since $p_2 \leq K/m$),

$$\begin{aligned}&\eta(k,n,p_1) \\ &\geq \min\{\eta(2np_1 - \sqrt{2n(1-p_1)p_1}, n, p_1), \eta(2np_1 + \sqrt{2n(1-p_1)p_1}, n, p_1)\} \\ &\geq \frac{1}{2\sqrt{2\pi}e^2\sqrt{n}}. \quad (2.2.23)\end{aligned}$$

Combining this last inequality with (2.2.22), and using Theorem 2.2.1, gives

$$\mathbb{P}\left(\left(\bigcap_{k\in I} A_n^k\right)^c\right) \leq 4\sqrt{\pi}e^2 n \mathbb{P}(A_n^c) \leq 816\sqrt{\pi}e^2 n \exp\left(-\frac{np_2^6}{5}\right). \quad (2.2.24)$$

Next, for each $n \geq 1$, let

$$\Delta_{k+1} = \begin{cases} LC_n(k+1) - LC_n(k), & \text{when } A_n^k \text{ holds,} \\ 1, & \text{otherwise,} \end{cases} \tag{2.2.25}$$

then from Theorem 2.2.1,

$$\mathbb{E}(\Delta_{k+1}|X^k, Y^k) \geq \frac{K_2}{2}. \tag{2.2.26}$$

For $k = 0, 1, \cdots, 2n$, let $\mathcal{F}_k := \sigma(X^0, Y^0, \cdots, X^k, Y^k)$, and let $V_k := \Delta_k - \mathbb{E}(\Delta_k|\mathcal{F}_{k-1})$, $k \geq 1$. Then $(V_k)_{1\leq k\leq 2n}$ forms a martingale differences with respect to $(\mathcal{F}_k)_{0\leq k\leq 2n-1}$, and since $-1 \leq \Delta_k \leq 1$, it follows from Hoeffding's martingale inequality that, for any $i < j$,

$$\mathbb{P}\left(\sum_{k=i+1}^{j} V_k < -\frac{K_2}{4}(j-i)\right) \leq \exp\left(-\frac{K_2^2(j-i)}{32}\right). \tag{2.2.27}$$

Moreover, from (2.2.26),

$$\sum_{k=i+1}^{j} \mathbb{E}(\Delta_k|X^{k-1}, Y^{k-1}) \geq \frac{K_2}{2}(j-i),$$

thus

$$\mathbb{P}\left(\sum_{k=i+1}^{j} \Delta_k < \frac{K_2}{4}(j-i)\right) \leq \mathbb{P}\left(\sum_{k=i+1}^{j} V_k < -\frac{K_2}{4}(j-i)\right) \leq \exp\left(-\frac{K_2^2(j-i)}{32}\right). \tag{2.2.28}$$

For each $n \geq 1$, let now

$$O_n^\Delta = \bigcap_{\substack{i,j\in I \\ i+\ell(n)\leq j}} \left\{\sum_{i+1}^{j} \Delta_k \geq \frac{K_2}{4}(j-i)\right\},$$

then, from (2.2.28)

$$\mathbb{P}\left(\left(O_n^\Delta\right)^c\right) \leq \sum_{\substack{i,j\in I \\ i+\ell(n)\leq j}} \mathbb{P}\left(\sum_{i+1}^{j} \Delta_k < \frac{K_2}{4}(j-i)\right) \leq 2n \exp\left(-\frac{K_2^2\ell(n)}{32}\right). \tag{2.2.29}$$

From the very definition of Δ_k in (2.2.25),

$$\bigcap_{k\in I} A_n^k \cap O_n^\Delta \subset O_n,$$

which implies that

$$\mathbb{P}\left((O_n)^c\right) \leq \mathbb{P}\left(\left(\bigcap_{k\in I} A_n^k\right)^c\right) + \mathbb{P}\left(\left(O_n^{\Delta}\right)^c\right)$$
$$\leq 816\sqrt{\pi}e^2 n \exp\left(-\frac{np_2^6}{5}\right) + 2n\exp\left(-\frac{K_2^2\ell(n)}{32}\right), \quad (2.2.30)$$

and this finishes the proof. ■

2.3 Proof of Theorem 2.2.1

2.3.1 Description of alignments

Let us begin with an example. Let $\mathcal{A}_3 = \{1, 2, 3\}$ and, say, let

$$X = 1213131112, \quad Y = 1113121112. \quad (2.3.1)$$

One optimal alignment corresponding to the longest common subsequence (LCS) of X and Y is

1	2		1	3	1	3		1	1	1	2
1		1	1	3	1		2	1	1	1	2

(2.3.2)

while another possible optimal alignment is

1	2	1		3	1	3		1	1	1	2
1		1	1	3	1		2	1	1	1	2

(2.3.3)

Comparing these two alignments, it is seen that the way the letters α_1, between aligned non-α_1 letters, are aligned is not important as long as a maximal number of such letters α_1 are aligned. Hence in general we will only describe which non-α_1 letters are aligned and assume that between pairs of aligned non-α_1 letters a maximal number of letters α_1 are aligned. In other words, we can take the two alignments (2.3.2) and (2.3.3) as the same alignment.

Call *cells*, the parts of the alignment between pairs of aligned non-α_1 letters. For

example, the alignment (2.3.2) can be decomposed into two cells $C(1)$ and $C(2)$ as

$$\begin{array}{cccccccccccc} \overbrace{}^{C(1),\ v_1=-1} & & & & & \overbrace{}^{C(2),\ v_2=0} & & & & & & \\ 1 & 2 & & 1 & 3 & 1 & 3 & & 1 & 1 & 1 & 2 \\ 1 & & 1 & 1 & 3 & 1 & & 2 & 1 & 1 & 1 & 2 \end{array} \tag{2.3.4}$$

where moreover, v_i denotes the difference between the number of letters α_1 in the X-part and the Y-part of the cell $C(i)$. Note that any alignment can be represented as a finite sequence of differences of the number of α_1 of its cells. For the alignment (2.3.2), this gives the representation $(v_1, v_2) = (-1, 0)$. The same X and Y might have different optimal alignments thus different representations. Above, for example, another optimal representation is via $(v_1, v_2) = (0, -1)$:

$$\begin{array}{cccccccccccc} \overbrace{}^{C(1),\ v_1=0} & & & & & & \overbrace{}^{C(2),\ v_2=-1} & & & & & \\ 1 & 2 & 1 & 3 & 1 & 3 & 1 & & 1 & 1 & & 2 \\ 1 & & 1 & & 1 & 3 & 1 & 2 & 1 & 1 & 1 & 2 \end{array} \tag{2.3.5}$$

Let $X = X_1X_2\cdots X_n$ and $Y = Y_1Y_2\cdots Y_n$ be given. As just explained, to every optimal alignment, corresponds a vector representation $v := (v_1, \cdots, v_k)$ showing the number of cells (k, here) in the alignment and the difference between the number of letters α_1 in the cells. In every cell, the maximum amount of letters α_1 is aligned. On the other hand, for every $v = (v_1, \cdots, v_k) \in \mathbb{Z}^k$ corresponds a (possible empty) family of alignments. All of these alignments have the same pairs of aligned non-α_1 letters and between consecutive pairs of aligned non-α_1 letters, a maximal number of letters α_1 are aligned. Since the alignments corresponding to the same v can only differ in the way the letters α_1 are aligned inside the cells, we identify the alignments in the family associated with v as a single alignment. In other words, we identify each vector v with an alignment and vice-versa.

Writing $|v|$ for the number of coordinates of v, *i.e.*, $|v| = k$ if $v \in \mathbb{Z}^k$, then we can define the alignment associated with $v = (v_1, \cdots, v_k) \in \mathbb{Z}^k$ rigorously.

Definition 2.3.1 *Let $k \in \mathbb{N}$ and let $v = (v_1, \cdots, v_k) \in \mathbb{Z}^k$. Define $\pi_v(i), \nu_v(i)$ by induction on i: starting with $\pi_v(0) = \nu_v(0) = 0$, for $0 \leq i < k$, once $\pi_v(i), \nu_v(i)$ is defined, let $(\pi_v(i+1), \nu_v(i+1))$ be the smallest (s,t) (where $(s_1, t_1) \leq (s_2, t_2)$ means $s_1 \leq s_2$ and $t_1 \leq t_2$) such that the following three conditions are satisfied.*

1. *$\pi_v(i) < s$ and $\nu_v(i) < t$;*
2. *$X_s = Y_t \in \{\alpha_2, \cdots, \alpha_m\}$;*
3. *The difference between the number of letters α_1 in the interval $[\pi_v(i), s]$ and $[\nu_v(i), t]$ is equal to v_{i+1}.*

If no such (s,t) exists, then $\pi_v(i+1) = \cdots = \pi_v(k) = \infty$ and $\nu_v(i+1) = \cdots = \nu_v(k) = \infty$. In other words, $\pi_v(i), \nu_v(i)$ are the indices corresponding to the i-th non-α_1 aligned pair in v. The i-th cell $C_v(i)$ is the pair of sequences is

$$C_v(i) := \left(\left(X_{\pi_v(i-1)+1}, \cdots, X_{\pi_v(i)} \right), \left(Y_{\nu_v(i-1)+1}, \cdots, Y_{\nu_v(i)} \right) \right),$$

and we call the cell $C_v(i)$ a v_i-cell.

We can then let the alignment v be any alignment (provided that there exists at least one) such that the following three conditions hold:

1. $X_{\pi_v(i)}$ is aligned with $Y_{\nu_v(i)}$, for every $i = 1, 2, \cdots, k$;
2. the number of aligned letters α_1 in the cell $C_v(i)$, denoted by $S_v(i)$, is the minimum number of letters α_1 present in either $X_{\pi_v(i-1)+1}, \cdots, X_{\pi_v(i)}$ or $Y_{\nu_v(i-1)+1}, \cdots, Y_{\nu_v(i)}$;
3. after aligning $X_{\pi_v(k)}$ with $Y_{\nu_v(k)}$, align as many α_1 as possible, and let that number be r_v.

From these definitions, for any $v \in \mathbb{Z}^k$, if an alignment corresponding to v exists, then $\pi_v(k) \leq n$ and $\nu_v(k) \leq n$, and such a vector v is called *admissible*. Let V denote

the set of all admissible alignments, that is,

$$V := \left\{ v \in \bigcup_{k>0} \mathbb{Z}^k : \pi_v(|v|), \nu_v(|v|) \leq n \right\}. \tag{2.3.6}$$

Then for every $v \in V$, the length of the common subsequence corresponding to this alignment v is:

$$\ell C_v = |v| + \sum_{i=1}^{|v|} S_v(i) + r_v. \tag{2.3.7}$$

Therefore the length of the longest common subsequence of X and Y can be expressed as:

$$LC_n = \max_{v \in V} \ell C_v, \tag{2.3.8}$$

and an admissible alignment is optimal if and only if $\ell C_v = LC_n$.

In our example (2.3.1), $v = (-1, 0)$ and $(0, -1)$ are two optimal alignments, while the alignment $v = (-3, 2)$ is not optimal:

$$\begin{array}{cccccc|cccccccc} \overbrace{}^{C(1),\ v_1=-3} & & & & & & \overbrace{}^{C(2),\ v_2=2} & & & & & & & \\ 1 & & & & & 2 & 1 & 3 & 1 & 3 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 3 & 1 & 2 & & & & & 1 & 1 & 1 & 2 \end{array} \tag{2.3.9}$$

2.3.2 The effect of changing a non-α_1 letter to α_1

Again, the main idea behind Theorem 2.2.1 is that, by changing a randomly picked non-α_1 letter into α_1, the length of the longest common subsequence LC_n is more likely to increase by one than to decrease by one. More precisely, conditional on the event $A_n = \{(X, Y) \in \mathcal{B}_n\}$, the probability of an increase of LC_n is at least K_2 while the probability of a decrease is at most $K_2/2$. Let us illustrate this fact with an example. Let X and Y be given by,

$$X = 112113112131,\ Y = 131111111131, \tag{2.3.10}$$

with optimal alignment:

$$\overbrace{\begin{array}{cccccccccccccccc} 1 & & & 1 & 2 & 1 & 1 & 3 & 1 & 1 & 2 & 1 & & & & 3 \\ 1 & 3 & 1 & & & 1 & 1 & & 1 & 1 & & 1 & 1 & 1 & 1 & 3 \end{array}}^{C(1),\ v_1=-2} \begin{array}{c} 1 \\ 1 \end{array} \tag{2.3.11}$$

Above, there are 6 non-α_1 letters, $X_3, X_6, X_9, X_{11}, Y_2, Y_{11}$, each one of them has a probability of 1/6 to be picked and replaced by α_1. X_3, X_6, X_9 and Y_2 are not aligned, and moreover, since X_3, X_6, X_9 are on the top strand which contains a lesser number of α_1, picking one of them and replacing it, leads to an increase of one in the length of the LCS. On the other hand, since X_{11} and Y_{11} are aligned in this optimal alignment, picking one of them could potentially (but not necessarily) decrease the length of the LCS by one. Picking Y_2 does not reduce the length of the LCS, but may potentially increase the LCS by modifying the optimal alignment. In conclusion, in this example, by switching a randomly chosen non-α_1 letter into α_1, the probability to get an increase of the LCS is at least 1/2 while the probability to get a decrease of the LCS is at most 1/3.

To prove Theorem 2.2.1, we just need to prove that typically there exists an optimal alignment v such that:

1. Among all the non-α_1 letters in X and Y, the proportion which are on the cell-strand with smaller number of letters α_1, is at least K_2.

2. Among all the non-α_1 letters in X and Y, the proportion which are aligned is at most $K_2/2$.

Rigorously, let $v = (v_1, \cdots, v_k) \in \mathbb{Z}^k$ be admissible, and let $N_v^-(i)$ be the number of non-α_1 letters on the cell-strand of $C_v(i)$ with less α_1:

$$N_v^-(i) = \begin{cases} 0, & \text{if } v_i = 0, \\ \sum_{j=\pi_v(i-1)+1}^{\pi_v(i)-1} \mathbf{1}_{X_j \in \{\alpha_2, \cdots, \alpha_m\}}, & \text{if } v_i < 0, \\ \sum_{j=\nu_v(i-1)+1}^{\nu_v(i)-1} \mathbf{1}_{Y_j \in \{\alpha_2, \cdots, \alpha_m\}}, & \text{if } v_i > 0. \end{cases} \tag{2.3.12}$$

Then the total number of non-α_1 letters on cell-strand with smaller number of letters α_1 is

$$N_v^- := \sum_{i=1}^{|v|} N_v^-(i). \tag{2.3.13}$$

Let N_i be the number of letters α_i in the two sequences X and Y, and let

$$N_{>1} = \sum_{i=2}^{m} N_i, \tag{2.3.14}$$

then the set $\mathcal{B}_n$ can be defined as the set of pairs of sequences $(x, y) \in \mathcal{A}_m^n \times \mathcal{A}_m^n$ such that there exists an optimal alignment v satisfying $N_v^- \geq K_2 N_{>1}$ and $2|v| \leq K_2 N_{>1}/2$. Clearly, $\mathcal{B}_n$ depends on K_2. Recall that $A_n = \{(X, Y) \in \mathcal{B}_n\}$, and our next job is to prove that there exists $K_2 > 0$, such that

$$\mathbb{P}(A_n) \geq 1 - e^{-C_1 n},$$

for some $C_1 > 0$, independent of n.

In this argument, we need an optimal alignment with enough ones in the cell-strands with less number of letters α_1, the problem is that many optimal alignments can have most cells are 0-cells, *i.e.*, with the same number of letters α_1 on both strands. To solve this problem, for an optimal alignment with most cells are 0-cells, we would break up some of the 0-cells to create enough nonzero-cells. Meanwhile, we will maintain the alignment after the operation to be an optimal alignment. Let us first look at an example. Take two sequences

$$X = 1121131123, \quad Y = 112131113,$$

one of their optimal alignment is

$$\begin{array}{cccccccccccc}
\overbrace{}^{C(1),\ v_1=0} & & & \overbrace{}^{C(2),\ v_2=0} & & & & & & & & \\
1 & 1 & 2 & 1 & & & 1 & 3 & 1 & 1 & 2 & 3 \\
1 & 1 & 2 & 1 & 3 & 1 & & & 1 & 1 & & 3
\end{array} \tag{2.3.15}$$

with both cells $C(1)$ and $C(2)$ are 0-cells. Now if we look at X_6 and Y_5 in cell $C(2)$, they are only one position away from being aligned. Thus if we align them, instead

of the pair of X_5 and Y_6, we would break the cell $C(2)$ into two new cells $\tilde{C}(2)$ and $\tilde{C}(3)$, with $\tilde{v}_2 = 1$ and $\tilde{v}_3 = -1$. The new alignment after the operation:

$$\begin{array}{cccccccccccc} \overbrace{}^{\tilde{C}(1),\ \tilde{v}_1=0} & & & \overbrace{}^{\tilde{C}(2),\ \tilde{v}_2=1} & & & \overbrace{}^{\tilde{C}(3),\ \tilde{v}_3=-1} & & & & & \\ 1 & 1 & 2 & 1 & 1 & 3 & 1 & 1 & & & 2 & 3 \\ 1 & 1 & 2 & 1 & & 3 & 1 & 1 & 1 & & & 3 \end{array} \tag{2.3.16}$$

The advantage of breaking up a cell is that the new cells will have different number of letters α_1 on each strand, thus N_v^- tends to increase in this process while the score remains the same. Once we apply this operation and get enough cells with different number of letters α_1 on two strands, we would have a high probability to find enough non-α_1 letters on strands with smaller number of letters α_1.

The previous example leads to our next definition.

Definition 2.3.2 *Let $k \in \mathbb{N}$, $v \in \mathbb{Z}^k \cap V$, $1 \leq i \leq k$ and $v_i = 0$. The cell $C_v(i)$ is said to be breakable if there exists j and j' such that:*

1. *$X_j = Y_{j'} \in \{\alpha_2, \cdots, \alpha_m\}$;*
2. *$\pi_v(i-1) < j < \pi_v(i)$ and $\nu_v(i-1) < j' < \nu_v(i)$;*
3. *The difference between the number of letters α_1 in*

$$X_{\pi_v(i-1)+1} X_{\pi_v(i-1)+2} \cdots X_{j-1} \text{ and } Y_{\nu_v(i-1)+1} Y_{\nu_v(i-1)+2} \cdots Y_{j'-1}$$

is plus or minus one.

2.3.3 Probabilistic developments

After the combinational analysis of the previous sections, let us now bring in the picture some probabilistic tools. We start by introducing a useful way of constructing alignments corresponding a given vector $v = (v_1, \cdots, v_k) \in \mathbb{R}^k$. For $1 \leq i \leq n$ and $2 \leq j \leq m$, let ξ_i^j be the number of α_j's between the $(i-1)$-th α_1 and the i-th α_1, with the convention that ξ_1^j is the number of α_j's before the first α_1. For example, when $m = 3$

the two sequences $(\xi_1^2, \xi_2^2, \xi_3^2, \xi_4^2, \xi_5^2, \xi_6^2) = (0, 1, 0, 0, 0, 1)$ and $(\xi_1^3, \xi_2^3, \xi_3^3, \xi_4^3, \xi_5^3, \xi_6^3,) = (0, 1, 1, 0, 0, 0)$ correspond to $X = 1231311121$. Similarly, we define the corresponding random variables $(\eta_i^j)_{j=2}^m$ for Y.

Recall from Definition 2.3.1 that in order to construct 0-cell, we use the stopping time T_0, where

$$T_0 = \min_{2 \le j \le m} \min\{i = 1, 2, \cdots : \xi_i^j \neq 0, \eta_i^j \neq 0\}. \tag{2.3.17}$$

Likewise, to get a $-u$-cell ($u > 0$), we use the stopping time

$$T_{-u} = \min_{2 \le j \le m} \min\{i = 1, 2, \cdots : \xi_i^j \neq 0, \eta_{i+u}^j \neq 0\}, \tag{2.3.18}$$

and for the u-cell,

$$T_u = \min_{2 \le j \le m} \min\{i = 1, 2, \cdots : \xi_{i+u}^j \neq 0, \eta_i^j \neq 0\}. \tag{2.3.19}$$

In other words, a cell with $v_i = u$ can be constructed in the following way: first set u letters α_1 on the X strand, then align consecutive pairs of α_1, until meeting the first pair of non-α_1 letters.

Now let us look at the distribution of ξ_i^j. First, let

$$\xi_i^{>1} = \sum_{j=2}^{m} \xi_i^j$$

be the total number of non-α_1 letter between the $(i-1)$-th α_1 and i-th α_1. Then, $\xi_i^{>1}$ has a geometric distribution:

$$\mathbb{P}(\xi_i^{>1} = k) = (1 - p_1)^k p_1,$$

for $k = 0, 1, 2, \cdots$, where $(1 - p_1) = p_2 + p_3 + \cdots + p_m$. Moreover, since conditional on $\xi_i^{>1}$, $(\xi_i^j)_{j=2}^m$ has a multinomial distribution,

$$\begin{aligned}
\mathbb{P}(\xi_i^j = k) &= \sum_{l=k}^{\infty} \mathbb{P}(\xi_i^j = k | \xi_i^{>1} = l) \mathbb{P}(\xi_i^{>1} = l) \\
&= \sum_{l=k}^{\infty} \binom{l}{k} \left(\frac{p_j}{(1-p_1)}\right)^k \left(\frac{(1-p_1)-p_j}{(1-p_1)}\right)^{l-k} (1-p_1)^l p_1 \\
&= \left(\frac{p_1}{p_1 + p_j}\right) \left(\frac{p_j}{p_1 + p_j}\right)^k,
\end{aligned} \tag{2.3.20}$$

for $k = 0, 1, 2, \cdots$. Thus,

$$\xi_i^j + 1 \sim G\left(\frac{p_1}{p_1 + p_j}\right),$$

where $G(p)$ denotes the geometric distribution with parameter p.

We start with providing a rough lower bound for the length of the LCS. First, aligning as many letters α_1 as possible in X and Y, would get approximately a common subsequence of length np_1, then aligning as many letters α_2 as possible without disturbing the already aligned α_1, would give an additional

$$\sum_{i=1}^{np_1} \min\{\xi_i^2, \eta_i^2\},$$

aligned α_2. Moreover, since ξ_i^2 and η_i^2 are independent geometric random variables,

$$\min\{\xi_i^2, \eta_i^2\} + 1 \sim G\left(1 - \left(\frac{p_2}{p_1 + p_2}\right)^2\right), \tag{2.3.21}$$

so on average the aligned α_2 contribute

$$np_1 \frac{p_2^2}{p_1(p_1 + 2p_2)} = \frac{1}{p_1 + 2p_2} np_2^2 \geq (1 - p_2) np_2^2.$$

This heuristic idea leads to the following rigorous lemma:

Lemma 2.3.1 *Let $p_1 > 1/2$, and let $E_1 := \{LC_n \geq np_1 + ((1-p_2)^3 - p_2)\, np_2^2\}$. Then,*

$$\mathbb{P}(E_1) \geq 1 - 4\exp(-2np_2^6) - \exp\left(n(p_2^3 + \log(1 - p_2^3))(p_1 - p_2^3)\right).$$

Proof. For $\delta > 0$, let

$$E_2^x(\delta) := \left\{\left|\sum_{i=1}^{n} \mathbf{1}_{X_i = \alpha_1} - np_1\right| \leq \delta n\right\},$$

$$E_2^y(\delta) := \left\{\left|\sum_{i=1}^{n} \mathbf{1}_{Y_i = \alpha_1} - np_1\right| \leq \delta n\right\},$$

and let

$$E_2(\delta) := E_2^x(\delta) \cap E_2^y(\delta).$$

Therefore on $E_2(\delta)$, at least $n_1(\delta) := n(p_1 - \delta)$ letters α_1 can be aligned.

Let $\zeta_i := \min\{\xi_i^2, \eta_i^2\}$, from (2.3.21) $\zeta_i + 1$ has a geometric distribution. Now, if $Q_1, \cdots, Q_n$ are iid geometric distributed random variables with parameter p, then for every $\beta < 1$,

$$\mathbb{P}\left(\sum_{i=1}^{n} Q_i \leq \frac{\beta}{p} n\right) \leq \exp\left(-(\beta - 1 - \log\beta)n\right). \tag{2.3.22}$$

This last inequality implies that for every $\beta < 1$,

$$\mathbb{P}\left(\sum_{i=1}^{n_1(\delta)} \zeta_i < \frac{\beta n_1(\delta)}{1 - (\frac{p_2}{p_1+p_2})^2} - n_1(\delta)\right) \leq e^{-C(\beta)n_1(\delta)}, \tag{2.3.23}$$

where $C(\beta) = \beta - 1 - \log\beta$. Let now

$$E_3(\beta, \delta) := \left\{\sum_{i=1}^{n_1(\delta)} \zeta_i \geq \frac{\beta n_1(\delta)}{1 - (\frac{p_2}{p_1+p_2})^2} - n_1(\delta)\right\}.$$

Choosing $\delta = p_2^3$ and $\beta = 1 - p_2^3$, and when $E_2(\delta)$ and $E_3(\beta, \delta)$ hold, it follows that

$$\begin{aligned}
LC_n &\geq \frac{\beta n_1(\delta)}{1 - (\frac{p_2}{p_1+p_2})^2} - n_1(\delta) + n_1(\delta) \\
&= \frac{n_1(\delta)}{1 - (\frac{p_2}{p_1+p_2})^2} - n_1(\delta) + n_1(\delta) - \frac{n_1(\delta)p_2^3}{1 - (\frac{p_2}{p_1+p_2})^2} \\
&= np_2^2 \frac{p_1 - p_2^3}{(p_1+p_2)^2 - p_2^2} + n(p_1 - p_2^3) - np_2^2 \frac{p_2(p_1 - p_2^3)}{1 - (\frac{p_2}{p_1+p_2})^2} \\
&\geq np_1 + \left((1-p_2)^3 - p_2\right) np_2^2.
\end{aligned}$$

By Hoeffding's inequality, for any $\delta > 0$,

$$\mathbb{P}\left((E_2^x(\delta))^c\right) \leq 2e^{-2n\delta^2}, \quad \mathbb{P}\left((E_2^y(\delta))^c\right) \leq 2e^{-2n\delta^2},$$

and since $E_2(p_2^3) \cap E_3(1 - p_2^3, p_2^3) \subset E_1$,

$$\mathbb{P}((E_1)^c) \leq 4\exp(-2np_2^6) + \exp\left(n(p_2^3 + \log(1 - p_2^3))(p_1 - p_2^3)\right).$$

■

Let N_1^x be the number of α_1 in X, let N_1^y be the number of α_1 in Y, and so $N_1 = N_1^x + N_1^y$ is the number of α_1 in X and Y. From the proof of the previous lemma, we see that when $E_2(p_2^3)$ holds, N_1 is upper bounded by $2n(p_1+p_2^3)$, therefore on E_1, this yields

$$LC_n \geq \frac{N_1}{2} - np_2^3 + \left((1-p_2)^3 - p_2\right) np_2^2 \geq \frac{N_1}{2} + \frac{1}{2}np_2^2,$$

provided $p_2 \leq 1/10$. Now assume that $v = (v_1, \cdots, v_k) \in \mathbb{R}^k$ is any optimal alignment, then

$$\frac{N_1}{2} + \frac{1}{2}np_2^2 \leq LC_n \leq \frac{N_1 - \sum_{i=1}^k |v_i|}{2} + k,$$

which further yields

$$\sum_{i=1}^k |v_i| \leq 2k \text{ and } np_2^2 \leq 2k.$$

In other words, for any optimal alignment v,

$$v \in V_n := V \bigcap \left(\bigcup_{2k \geq np_2^2} V(k) \right), \tag{2.3.24}$$

where

$$V(k) := \left\{ (v_1, v_2, \cdots, v_k) \in \mathbb{Z}^k : |v_1| + \cdots + |v_k| \leq 2k \right\}, \tag{2.3.25}$$

and where V is the set of admissible alignments (see (2.3.6)). This leads to the fact that proving a property of the optimal alignment just requires proving it for every alignment in V_n. In conclusion, we have

Corollary 2.3.1 *Let $E = \{v \in V_n : v$ encodes an optimal alignment$\}$, and let $p_2 \leq 1/10$, then*

$$\mathbb{P}(E) \geq 1 - 5\exp\left(-\frac{np_2^6}{5}\right).$$

2.3.4 Events

Recall from Definition 2.3.1 that to a vector $v \in \mathbb{Z}^k$ is associated an alignment which has $|v|$ cells $C_v(1), \cdots, C_v(|v|)$. Such a cell is called a nonzero-cell if it contains a

different number of α_1 on the X strand and Y strand. Let V_1 be the subset of V_n, consisting of the alignments for which the proportion of the nonzero-cells is at least θ, *i.e.*,

$$W := \left\{ v \in V_n \middle| |\{i \in [1,k] | v_i \neq 0\}| \geq \theta k \right\},$$

and let

$$W^c := V_n - W.$$

Recall also that from (2.3.13) and (2.3.14), N_v^- is the number of non-α_1 letters on the cell strands with less α_1, and that $N_{>1}$ is the total number of non-α_1 letters in X and Y. Next we will define some events.

- Let D be the event that for all $v \in W^c$, the proportion of the breakable zero-cells is at least θ, *i.e.*,

  $$D := \bigcap_{v \in W^c} D_v,$$

  where D_v is the event that, among the zero-cells in $C_v(1), \cdots, C_v(|v|)$, the proportion of the them which can be breakable is at least θ.

- Let

  $$F := \bigcap_{v \in W} F_v := \bigcap_{v \in W} \{N_v^- \geq K_2 N_{>1}\},$$

  in other words, F is the event that for every $v \in W$, among all the non-α_1 letters, the proportion which are on the cell-strands with smaller number of α_1 letters, is at least K_2.

- Let

  $$H := \bigcap_{v \in W} H_v := \bigcap_{v \in W} \{2|v| \leq \frac{K_2}{2} N_{>1}\},$$

  in other words, H is the event that for every $v \in W$, among all the non-α_1 letters, the proportion which are aligned is at most $K_2/2$.

Recall from Section 2.3.2 that $A_n = \{(X, Y) \in \mathcal{B}_n\}$ is the event that there exists an optimal alignment v such that $N_v^- \geq K_2 N_{>1}$ and $2|v| \leq K_2 N_{>1}/2$, thus

$$E \cap D \cap F \cap H \subset A_n, \tag{2.3.26}$$

thus our next job is to prove that there exists $K_2 > 0$ such that the events D, F, H hold with high probability.

Lemma 2.3.2 *For any* $0 < \theta < 1$,

$$\mathbb{P}(D) \geq 1 - \sum_{2k \geq np_2^2} \exp\left(- \left(2(1-\theta) \left(\frac{p_1^2}{1+p_1^2} - \theta \right)^2 - \frac{3}{20} \right) k \right). \tag{2.3.27}$$

Proof. For any $v \in W^c$, let us calculate the probability that a 0-cell in the alignment associated with v is breakable. Recall from the definition of T_0 in (2.3.17), and for $2 \leq j \leq m$, let M_j be the event this cell ends with a pair of α_j's. Define

$$T_0^j := \min\{i = 1, 2, \cdots : \quad \xi_i^j \neq 0, \eta_i^j \neq 0\}, \tag{2.3.28}$$

thus $T_0 = \min_{2 \leq j \leq m} T_0^j$ and when M_j holds $T_0 = T_0^j$.

For $2 \leq j \leq m$, let

$$U_1^j := \min\{i = 2, \cdots : \quad \xi_{i-1}^j \neq 0, \quad \eta_{i-1}^j = 0, \quad \xi_i^j = 0, \quad \eta_i^j \neq 0\},$$

$$U_2^j := \min\{i = 2, \cdots : \quad \xi_{i-1}^j = 0, \quad \eta_{i-1}^j \neq 0, \quad \xi_i^j \neq 0, \quad \eta_i^j = 0\},$$

$$U^j := \min\{U_1^j, U_2^j\}.$$

With the above constructions, conditional on the event M_j, $U^j < T_0^j$ implies that this 0-cell is breakable, thus we need estimate $\mathbb{P}(U^j < T_0^j)$ from below. To do this, for $i = 1, 2, \cdots$, define the independent random vectors

$$Z_i^j = (\xi_{2i-1}^j, \eta_{2i-1}^j, \xi_{2i}^j, \eta_{2i}^j),$$

$$\tilde{U}^j = \min\{i = 1, 2, \cdots : \quad Z_i^j \in A_1 \cup A_2\},$$

$$\tilde{T}_0^j = \min\{i = 1, 2, \cdots : \quad Z_i^j \in B_1 \cup B_2\},$$

where

$$A_1 := \mathbb{N}^* \times \{0\} \times \{0\} \times \mathbb{N}^*, A_2 = \{0\} \times \mathbb{N}^* \times \mathbb{N}^* \times \{0\},$$

$$B_1 := \mathbb{N}^* \times \mathbb{N}^* \times \mathbb{N} \times \mathbb{N}, \quad B_2 := \mathbb{N} \times \mathbb{N} \times \mathbb{N}^* \times \mathbb{N}^*,$$

and where as usual $\mathbb{N}$ is the set of nonnegative integers while $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$. Clearly,

$$2\tilde{U}^j \geq U^j, \quad 2\tilde{T}_0^j - 1 \leq T_0^j,$$

thus

$$\mathbb{P}(U^j < T_0^j) \geq \mathbb{P}(2\tilde{U}^j < 2\tilde{T}_0^j - 1) = \mathbb{P}(\tilde{U}^j < \tilde{T}_0^j).$$

Since the random variables $(Z_i^j)_{i \in \mathbb{N}^+}$ are iid, and since $A_1 \cup A_2$ and $B_1 \cup B_2$ are pairwise disjoint,

$$\begin{aligned}
\mathbb{P}(\tilde{U}^j < \tilde{T}_0^j) &= \frac{\mathbb{P}(Z_i^j \in A_1 \cup A_2)}{\mathbb{P}(Z_i^j \in A_1 \cup A_2) + \mathbb{P}(Z_i^j \in B_1 \cup B_2)} \\
&= \frac{2p_1^2}{2p_1^2 + 2(p_1 + p_j)^2 - p_j^2} \\
&\geq \frac{p_1^2}{1 + p_1^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\text{0-cell is breakable}) &= \sum_{j=2}^{m} \mathbb{P}(\text{0-cell is breakable}|M_j)\mathbb{P}(M_j) \\
&\geq \sum_{j=2}^{m} \mathbb{P}(U^j < T_0^j|M_j)\mathbb{P}(M_j) \\
&\geq \frac{p_1^2}{1 + p_1^2}.
\end{aligned}$$

Let J be the index set of 0-cells in the alignment associated with $v \in W^c$, thus $|J| \geq (1 - \theta)|v|$. For every $i \in J$, let I_i be the Bernoulli variable that is one if and only if the cell $C_v(i)$ is breakable. Recall that D_v is the event that the proportion of

the cells that can be breakable in v is at least θ, then from Hoeffding's inequality,

$$\begin{aligned}\mathbb{P}(D_v^c) &= \mathbb{P}\left(\sum_{i\in J} I_i < \theta|J|\right)\\ &= \mathbb{P}\left(\sum_{i\in J} I_i - \mathbb{E}\left(\sum_{i\in J} I_i\right) < \theta|J| - \mathbb{E}\left(\sum_{i\in J} I_i\right)\right)\\ &\le \exp\left(-2(1-\theta)|v|\left(\frac{p_1^2}{1+p_1^2}-\theta\right)^2\right).\end{aligned}$$

Recall the definition of $V(k)$ in (2.3.25), and let $W^c(k) := W^c \cap V(k)$, proceeding from [30], $|W^c(k)| \le \exp(3k/20)$. Let

$$D(k) = \bigcap_{v\in W^c(k)} D_v,$$

then

$$\mathbb{P}(D^c(k)) \le \sum_{v\in W^c(k)} \mathbb{P}(D_v^c) \le \exp\left(-\left(2(1-\theta)\left(\frac{p_1^2}{1+p_1^2}-\theta\right)^2 - \frac{3}{20}\right)k\right),$$

which further gives

$$\mathbb{P}(D^c) \le \sum_{2k\ge np_2^2} \mathbb{P}(D^c(k)) \le \sum_{2k\ge np_2^2} \exp\left(-\left(2(1-\theta)\left(\frac{p_1^2}{1+p_1^2}-\theta\right)^2 - \frac{3}{20}\right)k\right).$$

■

Let u be a nonnegative integer. For any $-u$-cell ending with an aligned pair of α_j (*i.e.* K_j holds for this cell), let $\tau_x^j(l)$ be the index of l-th ξ_i^j such that $\xi_i^j \neq 0$, in other words,

$$\tau_x^j(1) = \min\{i \ge 1 :\ \xi_i^j \neq 0\},$$

and for any $l \ge 1$, $\tau_x^j(l+1) = \min\{i > \tau_x^j(l) :\ \xi_i^j \neq 0\}$.

Let

$$\rho^{j,-} := \min\{l = 1, 2, \cdots :\ \eta^j_{u+\tau_x^j(l)} \neq 0\}.$$

Hence $\rho^{j,-}$ is the number of nonzero ξ_i^j's in the cell (including the last one corresponding to the aligned pair of α_j). Since X and Y are independent, $\rho^{j,-}$ has a geometric

distribution

$$\begin{aligned}\mathbb{P}(\rho^{j,-}=k) &= \mathbb{P}(\eta^j_{u+\tau^j_x(1)}=0,\cdots,\eta^j_{u+\tau^j_x(k-1)}=0,\eta^j_{u+\tau^j_x(k)}\neq 0)\\ &= \left(\frac{p_1}{p_1+p_j}\right)^{k-1}\frac{p_j}{p_1+p_j},\end{aligned} \tag{2.3.29}$$

for $k=1,2,\cdots$. Let $\tilde{p}_j = p_j/(p_1+p_j)$, then $\rho^{j,-}\sim G(\tilde{p}_j)$, for $2\leq j\leq m$. When $-u<0$, the number of letters α_j on the X-strand (strand with smaller α_1) is at least $\rho^{j,-}-1$, and this gives a lower bound for N_v^- (the number of non-α_1 letters on the cell-strand with less number of letters α_1) in this $-u$-cell.

Lemma 2.3.3 *Let $K_2 = K_1/m$, where $K_1 = 2^{-4}10^{-1}e^{-16}$ and where m is the alphabet size, then for any $p_1 \geq 1-2^{-3}e^{-16}$,*

$$\mathbb{P}(F)\geq 1-15\exp(-np_2^2/20).$$

Proof. For any $v\in W$, let now J be the index set of nonzero cells of the alignment corresponding to v, thus $|J|\geq\theta|v|$.

$$N_v^- = \sum_{i=1}^{|v|}N_v^-(i) = \sum_{i\in J}N_v^-(i)\geq\sum_{i\in J}\left(\rho_i^{j(i),-}-1\right),$$

where $j(i)$ is the index of the last aligned pair α_j in cell $C_v(i)$, and $\rho_i^{j(i),-}$ is the number of $\xi^{j(i)}(l)$ (or $\eta^{j(i)}(l)$, depending on which cell strand has less number of letters α_1) in the cell $C_v(i)$, where $\xi^{j(i)}(l)$ is the l-th nonzero $\xi^{j(i)}$ in the X-strand of $C_v(i)$. From (2.3.29), $\rho_i^{j(i),-}\sim G(\tilde{p}_{j(i)})$. Now let γ be some positive constant, and let

$$F_{1v} := \left\{N_v^-\geq\frac{\gamma}{\tilde{p}_2}|v|\right\},$$

then,

$$\begin{aligned}\mathbb{P}(F_{1v}^c) &\leq \mathbb{P}\left(\sum_{i\in J}\left(\rho_i^{j(i),-}-1\right)\leq \frac{\gamma}{\tilde{p}_2}|v|\right)\\ &= \mathbb{P}\left(\sum_{i\in J}\left(\rho_i^{j(i),-}\right)\leq \frac{\gamma}{\tilde{p}_2}|v|+|J|\right)\\ &\leq \mathbb{P}\left(\sum_{i\in J}\left(\rho_i^{j(i),-}\right)\leq \frac{\gamma/\theta+\tilde{p}_2}{\tilde{p}_2}|J|\right)\\ &\leq \mathbb{P}\left(\sum_{i\in J}\left(\rho_i^{j(i),-}\right)\leq \frac{\gamma/\theta+2p_2}{\tilde{p}_2}|J|\right). \end{aligned} \tag{2.3.30}$$

Since for $i\in J$, the geometric random variables $\rho_i^{j(i),-}$ are independent with parameters $\tilde{p}_{j(i)}\leq \tilde{p}_2$, it follows that

$$\mathbb{P}\left(\sum_{i\in J}\left(\rho_i^{j(i),-}\right)\leq \frac{\gamma/\theta+2p_2}{\tilde{p}_2}|J|\right)\leq \mathbb{P}\left(\sum_{i\in J}G_i\leq \frac{\gamma/\theta+2p_2}{\tilde{p}_2}|J|\right), \tag{2.3.31}$$

where G_i are iid geometric random variables with parameter $\tilde{p}_2$.

From (2.3.22),

$$\mathbb{P}\left(\sum_{i\in J}G_i\leq \frac{\gamma/\theta+2p_2}{\tilde{p}_2}|J|\right)\leq \exp\left(-\left(-1-\log(\gamma/\theta+2p_2)\right)\theta|v|\right), \tag{2.3.32}$$

moreover, let

$$F_1(k):=\bigcap_{v\in W\cap V(k)}F_{1v} \text{ and } F_1:=\bigcap_{2k\geq np_2^2}F_1(k),$$

since from [30], $|V(k)|\leq 16^k$, thus

$$\begin{aligned}\mathbb{P}(F_1(k)^c) &\leq 16^k\exp\left(-\left(-1-\log(\gamma/\theta+2p_2)\right)\theta k\right)\\ &= \exp\left(k\log 16-\left(-1-\log(\gamma/\theta+2p_2)\right)\theta k\right). \end{aligned} \tag{2.3.33}$$

Since we need have $\mathbb{P}(F_1(k)^c)$ going to 0 as $k\to\infty$, thus we need

$$\log 16-\left(-1-\log(\gamma/\theta+2p_2)\right)\theta<0. \tag{2.3.34}$$

For similar reasons, from (2.3.27), we need to guarantee that

$$2(1-\theta)\left(\frac{p_1^2}{1+p_1^2}-\theta\right)^2-\frac{3}{20}>0. \tag{2.3.35}$$

Combining (2.3.34) and (2.3.35), by choosing $\theta = 19/100$, $\gamma = 10^{-1}e^{-16}$, then for any $p_1 \geq 1 - 2^{-3}e^{-16}$,

$$\mathbb{P}(F_1(k)^c) \leq e^{-k/10},$$

and so

$$\mathbb{P}(F_1^c) \leq \sum_{2k \geq np_2^2} \mathbb{P}(F_1(k)^c) \leq 11 \exp(-np_2^2/20).$$

Recall from the proof of Lemma 2.3.1 that, when $E_2((1-p_1))$ holds, the total number of non-α_1 letters in X and Y is at most $4n(1-p_1)$. Thus $N_{>1} \leq 4n(1-p_1)$, and so when $F_1 \cap E_2((1-p_1))$ holds, for every $v \in W$,

$$\frac{N_v^-}{N_{>1}} \geq \frac{\gamma|v|}{\tilde{p}_2 4n(1-p_1)} \geq \frac{\gamma}{\tilde{p}_2 4n(1-p_1)} \frac{np_2^2}{2} \geq \frac{\gamma p_2}{16(1-p_1)} \geq \frac{\gamma}{16m} = K_2,$$

which implies F. Therefore,

$$\begin{aligned}\mathbb{P}(F^c) \leq \mathbb{P}(F_1^c) + \mathbb{P}((E_2(1-p_1))^c) &\leq 11\exp(-np_2^2/20) + 4\exp(-2n(1-p_1)^2) \\ &\leq 15\exp(-np_2^2/20).\end{aligned}$$

■

Lemma 2.3.4 *Let $K_2 = K_1/m$, then for any $p_2 \leq 2^{-2}e^{-5}K_2$*

$$\mathbb{P}(H) \geq 1 - 4\exp(-np_2^2/2).$$

Proof. For any $v \in W$, let $C_v(1), \cdots, C_v(|v|)$ be the corresponding cells. For the cell $C_v(i)$, if it is ending with a pair of aligned α_j, $2 \leq j \leq m$, then let $\rho_i^j(i)$ be the number of nonzero $\xi^j(l)$'s in $C_v(i)$. If $v_i \leq 0$, from the same argument as in (2.3.29), $\rho_i^j(i)$ has a geometric distribution with parameter $\tilde{p}_{j(i)}$. If $v_i > 0$, there exists a geometric random variable $\rho_i^{j(i),-}$ with parameter $\tilde{p}_{j(i)}$ such that $\rho_i^{j(i),-} \leq \rho_i^{j(i)} \leq \rho_i^{j(i),-} + v_i$. Let $N_{>1}^x$ be the number of non-α_1 letters in X and $N_{>1}^y$ be the number of non-α_1 letters in Y, so that $N_{>1} = N_{>1}^x + N_{>1}^y$. Let

$$H_v^x := \left\{|v| \leq \frac{K_2}{2} N_{>1}^x\right\} \text{ and } H_v^y := \left\{|v| \leq \frac{K_2}{2} N_{>1}^y\right\},$$

thus

$$H_v^x \cap H_v^y \subset H_v.$$

Since

$$N_{>1}^x \geq \sum_{i=1}^{|v|} \rho_i^{j(i)},$$

thus when $p_2 \leq 2^{-2}e^{-5}K_2$,

$$\begin{aligned}
\mathbb{P}\left((H_v^x)^c\right) &\leq \mathbb{P}\left(|v| > \frac{K_2}{2}\sum_{i=1}^{|v|}\rho_i^{j(i)}\right) \\
&\leq \mathbb{P}\left(|v| > \frac{K_2}{2}\left(\sum_{1\leq i\leq|v|,v_i\leq 0}\rho_i^{j(i)} + \sum_{1\leq i\leq|v|,v_i>0}\rho_i^{j(i),-}\right)\right) \\
&\leq \mathbb{P}\left(|v| > \frac{K_2}{2}\sum_{i=1}^{|v|}G_i\right) \\
&\leq \mathbb{P}\left(\sum_{i=1}^{|v|}G_i < \frac{e^{-5}|v|}{\tilde{p}_2}\right) \\
&\leq \exp(-4|v|),
\end{aligned}$$

where the G_i are iid geometric random variable with parameter $\tilde{p}_2$. Likewise we prove that

$$\mathbb{P}\left((H_v^y)^c\right) \leq \exp(-4|v|),$$

and thus

$$\mathbb{P}\left((H_v)^c\right) \leq 2\exp(-4|v|).$$

As previously, let

$$H(k) := \bigcap_{v\in W\cap V(k)} H_v \text{ and thus } H = \bigcap_{2k\geq np_2^2} H(k),$$

$$\mathbb{P}((H(k))^c) \leq |V(k)|2\exp(-4k) \leq 2\exp(-k),$$

and

$$\mathbb{P}(H^c) \leq \sum_{2k\geq np_2^2}\mathbb{P}((H(k))^c) \leq 4\exp(-np_2^2/2).$$

■

Combining Corollary 2.3.1, Lemma 2.3.2, 2.3.3, 2.3.4, using (2.3.26), letting $K_2 = K_1/m$, and $\theta = 19/100$, it follows that for $p_2 \leq 2^{-2}e^{-5}K_2 = K/m$,

$$\begin{aligned}
&\mathbb{P}(A_n^c) \leq \mathbb{P}(E^c) + \mathbb{P}(D^c) + \mathbb{P}(F^c) + \mathbb{P}(H^c) \\
&\leq 5\exp\left(-\frac{np_2^6}{5}\right) + 180e^{-5^2 10^{-4} np_2^2} + 15\exp(-np_2^2/20) + 4\exp(-np_2^2/2) \\
&\leq 204\exp\left(-\frac{np_2^6}{5}\right), \qquad (2.3.36)
\end{aligned}$$

and this finishes the proof of Theorem 2.2.1.

CHAPTER III

CONCLUSION

In this thesis, we have tackled two sets of problems in sequence analysis. In the first set, for the uniform model, we first investigate the large deviations of the eigenvalues of the traceless GUE. When the eigenvalues are on the right of the asymptotic mean, since the distribution of the spectrum of the GUE and of the traceless GUE only differ by a normal random vector, which becomes negligible in this case, they share the same large deviations principle. This is shown in Section 1.5.

When the eigenvalues are on the left of the asymptotic mean, the situation is different, and the normal vector is no longer negligible. The LDP has a different speed in this case, which is essentially due to the asymmetry of the Tracy-Widom distribution. Then, to get the rate function of the eigenvalues of the traceless GUE, two different approaches are used. The first approach observes that the empirical mean measure of the spectrum satisfies a LDP on the space of zero mean probability measures. This leads to a closed form expression for the rate function, using a calculus of variation argument. The second approach, coming from the relationship between the Legendre transform of the rate functions of the eigenvalues of the GUE and the traceless GUE, is also applicable to the non-uniform model.

When the large deviations of the eigenvalues of the traceless GUE are available, we use the KMT approximation to derive the large deviations for the shape of the random RSK Young diagrams. Indeed, this shape is sharing the same functional structure as the spectrum of the traceless GUE. This argument requires a control of the size of the alphabet when compared to the length of the random word, and it is contained in Section 1.2.

The non-uniform model is similar to the uniform model, but then the shape of the Young diagrams is related to the eigenvalues of the "generalized" traceless GUE, which depends on the probability distribution of the alphabet set.

In Section 1.4, non-asymptotic concentration bounds for the length of the top row of the diagrams are given, as a complement to the LDP results. In both the uniform and non-uniform model, the orders of the exponential decay in the concentrations are compatible with that in the LDP.

In the longest common subsequence part of the thesis, we proved a lower bound on the order of the central moment of the LCS, when all but one of the letters are drawn with small probabilities. This lower bound order matches the order of a known generic upper bound.

REFERENCES

[1] K. S. Alexander. *The rate of convergence of the mean length of the longest common subsequence.* Ann. Appl. Probab., 4(4), 1074-1082, 1994.

[2] K. S. Alexander. *Approximation of subadditive functions and convergence rates in limiting-shape results.* Ann. Probab., 25(1):30-55, 1997.

[3] G. Aubrun. *An inequality about the largest eigenvalue of a random matrix.* Séminaire de Probabilités XXXVIII 320-337. Lecture Notes in Math. 1857. Springer, Berlin, 2005.

[4] A. Auffinger, G. Ben Arous and J. Cerny. *Random matrices and complexity of spin glasses.* ArXiv: 1003.1129, 2010.

[5] G. W. Anderson, A. Guionnet and O. Zeitouni. *An Introduction to random matrices.* Cambridge University Press, 2010.

[6] S. Amsalu, C. Houdré, H. Matzinger. *Sparse long blocks and the variance of the LCS.* Preprint arXiv:1204.1009v1, 2012.

[7] J. Baik, T. Suidan. *A GUE central limit theorem and universality of directed first and last passage percolation site.* Int. Math. Res. Not. no. 6, pp. 325-337, 2005.

[8] Y. Baryshnikov. *GUEs and Queues.* Probab. Theor. and Relat. Fields, vol. 119, pp. 256-274, 2001.

[9] G. Ben Arous, A. Dembo and A. Guionnet. *Aging of spherical spin glasses.* Probab. Theory Related Fields 120, no. 1, 1-67, 2001.

[10] G. Ben Arous, A. Guionnet. *Large deviations for Wigner's law and Voiculescu's non-commutative entropy.* Probab. Theory Relat. Fields, 108, 517-542, 1997.

[11] J.-C. Breton, C. Houdré. *Asymptotics for random Young diagrams when the word length and alphabet size simultaneously grow to infinity.* Bernoulli. 16, 471-492, 2010.

[12] T. Bodineau, J. Martin. *A universality property for last-passage percolation paths close to the axis.* Elect. Comm. Probab. vol. 10, pp. 105-112, 2005.

[13] F. Bonetto, H. Matzinger. *Fluctuations of the longest common subsequence in the asymmetric case of 2- and 3-letter alphabets.* Alea, 2:195-216, 2006.

[14] V. Chvátal, D. Sankoff. *Longest common subsequences of two random sequences.* J. Appl. Probab. 12, 306-315, 1975.

[15] D. S. Dean, S. N. Majumdar. *Large deviations of extreme eigenvalues of random matrices.* Phys. Rev. Lett. 97, no. 16, 160210, 4pp, 2006.

[16] A. Dembo, O. Zeitouni. *Large deviations techniques and applications.* 2nd ed. New York , Springer, 1998.

[17] J.-D. Deuschel, O. Zeitouni. *On increasing subsequences of I.I.D. samples.* Combin. Probab. Comput. 8(3), 247-263, 1999.

[18] J. Gravner, C. Tracy and H. Widom. *Limit theorems for height fluctuations in a class of discrete space and time growth models.* J. Stat. Phys., vol. 102 nos 5-6, pp. 1085-1132, 2001.

[19] C. Houdré, T. Litherland. *On the longest increasing subsequence for finite and countable alphabets.* High Dimensional Probability V: The Luminy Volume, IMS Collections 185-212, 2009.

[20] C. Houdré, T. Litherland. *On the limiting shape of Young diagrams associated with Markov random words.* Preprint arXiv: 1110.4570, 2011.

[21] C. Houdré, H. Matzinger. *On the variance of the optimal alignment score for an asymmetric scoring function.* Preprint arXiv:math/0702036, 2007.

[22] C. Houdré, H. Xu. *On the limiting shape of random Young tableaux associated to inhomogeneous words.* To appear on: High Dimensional Probability VI: The Banff Volume, Springer, 2013.

[23] J-P. Ibrahim. *Large deviations for directed percolation on a thin rectangle.* Accepted at ESAIM P&S.

[24] K. Johansson. *Shape fluctuation and random matrices.* Comm. Math. Phys. 209, 437-476, 2000.

[25] K. Johansson. *Discrete polynomials ensembles and the Plancherel measure.* Ann. Math. vol. 153, pp. 259-296, 2001.

[26] J. Lember, H. Matzinger. *Standard deviation of the longest common subsequence.* Ann. Probab. 37, no. 3, 1192-1235, 2009.

[27] M. Ledoux, B. Rider. *Small deviations for beta ensembles.* Electron. J. Probab. 15, no. 41, 1319-1343, 2010.

[28] M. Lifshits. *Lecture notes on strong approximation.* Pub. IRMA Lille 53 13, 2000.

[29] M. Lifshits. *Gaussian Random Functions.* Kluwer Academic Publishers, 1995.

[30] M. Löwe, F. Merkl. *Moderate deviations for longest increasing subsequences: the upper tail.* Comm. Pure Appl. Math. 54, no. 12, 1488-1520, 2001.

[31] M. Löwe, F. Merkl and S. Rolles. *Moderate deviations for longest increasing subsequences: the lower tail.* J. Theoret. Probab. 15, no. 4, 1031–1047, 2002.

[32] M. L. Mehta. *Random matrices. 3nd ed.* Elsevier/Academic Press, Amsterdam, 2004.

[33] C. Nadal, S. N. Majumdar, M. Vergassola. *Statistical distribution of quantum entanglement for a random bipartite state.* J. Stat. Phys. 142, no.2, 403-438, 2011.

[34] R. T. Rockafellar. *Convex analysis.* Princeton University Press, Princeton, NJ, 1970.

[35] E. B. Saff, V. Totik. *Logarithmic potentials with external fields.* Springer, Berlin, 1997.

[36] T. Seppäläinen. *Large deviations for increasing sequences on the plane.* Probab. Theory Related Fields 112(2), 221-244, 1998.

[37] J. M. Steele. *An Efron-Stein inequality for nonsymmetric statistics.* Ann. Statist. 14, 753-758, 1986.

[38] C. Tracy, H. Widom. *Level-spacing distributions and the Airy kernel.* Commun. Math. Phys., vol. 159, pp. 151-174, 1994.

[39] C. Tracy, H. Widom. *On the distributions of the lengths of the longest monotone subsequences in random words.* Probab. Theory Relat. Fields, vol. 119, pp. 350-380, 2001.

[40] C. Tracy, H. Widom. *Matrix kernels for the Gaussian orthogonal and symplectic ensembles.* Annales de l'institut Fourier, 55 no. 6, p. 2197-2207, 2005.

[41] F. G. Tricomi. *Integral equations.* Pure Appl. Math., vol. V. Interscience, London, 1957.

[42] M. S. Waterman. *Estimating statistical significance of sequence alignments.* Philos. Trans. R. Soc. Lond. Ser. B 344, 383-390, 1994.

VITA

Jinyong Ma was born in Weifang of Shandong Province, China on February 10, 1985. He went to University of Science and Technology of China in Hefei in 2003 and graduated with a Bachelor degree of Science in mathematics in June of 2007. Following the completion of his Bachelor degree, he came to Atlanta in August and joined the Georgia Institute of Technology to work in Probability Theory with Professor Christian Houdré in the School of Mathematics. During his PhD study at Georgia Institute of Technology, he also received a secondary Masters degree of Science in Statistics from the School of Industrial and System Engineering in May of 2012.